1994

On Information Usage Modeling.

Ping Pete Chong
Louisiana State University and Agricultural & Mechanical College

Order Number 9502104

On information usage modeling

Chong, Ping Pete, Ph.D.

The Louisiana State University and Agricultural and Mechanical Col., 1994

U·M·I
300 N. Zeeb Rd.
Ann Arbor, MI 48106

ON INFORMATION USAGE MODELING

A Dissertation

Submitted to the Graduate Faculty of the
Louisiana State University and
Agricultural and Mechanical College
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

in

The Interdepartmental Program in Business Administration

by
Ping Pete Chong
B.A., Southeastern Louisiana University, 1985
A.S., Southeastern Louisiana University, 1986
M.B.A., Southeastern Louisiana University, 1987
May 1994

ACKNOWLEDGEMENTS

Above all, I owe my thanks to the maker who took me this far. Then I must thank Dr. Y. S. Chen, not only for his scholastic insight and know-how, but also for his genuine interest in helping his students achieving their goals. Without his consistent help over the years, it would have been impossible for me to reach this point. I will teach my students according to the examples set in having me taught.

My thanks to the committee members who gave me guidances and encouragements throughout the process. Though the frustration of dissertation makes it three times the four-letter word, I thank them for pointing out the right directions to go and the traps to avoid. I also thank them for concocting up ways to trick me into taking the right road, for at the end I turned out to be the greatest beneficiary.

My thanks to my friends: to SLU library staffs who assisted me in interpreting the data; to Kevin Calmes who helped generously in extracting the SLU data; to Dan and Fran who helped proof reading; and most of all, to Yueguo Tong who provided so much technical assistance in putting the autoregressive model together.

My thanks to my family: to my father who set an example of getting his Ph.D. at age fifty; to my in-laws who chipped in during difficult times; to my children who told me "Dad, we can't live on like this;" and to my dear dear wife for her putting up with the imperfect me. For her, I would give up my life.

This is a small step for mankind, but a giant step for me; so help me God.

TABLE OF CONTENTS

ABSTRACT

Although it is well-known that the usage of information usually follows the 80/20 rule and concentrates on a few items, there has not been an analytical model to depict this skew distribution. This dissertation provides a theoretical foundation, based on Simon's modeling of empirical phenomena and Chen's index approach, to identify the factors which shape this usage pattern. Using Chen's index approach, we conclude that the distance and slope of the data points determine the shape of the distribution. We further examine the critical parameters in Simon's model through computer simulations, and we find the probability of new entry (α) and the rate of "decay" (ß) to be two predominant factors that affect the patterns of information usage. Based on the effects of these two parameters we can establish the limiting conditions under which these empirical phenomena hold true. Finally, we show how our findings can be applied to enhance the weeding process in libraries — a procedure that can be extended to the archive management of information systems.

CHAPTER 1

INTRODUCTION

Usages in information systems (IS) usually follow a skew distribution: for examples, in a software some functions are used more than the others; and in the information retrieval process a relatively large number of activities are concentrated on a few records. In order to design more efficient IS based on these usage patterns, we need to: (1) identify factors that affect usages, and (2) identify the limiting conditions of our model. Fortunately, similar phenomena have been observed in social aggregates such as personal wealth, incomes, size of business firms or cities, publication frequencies, and word frequencies (Ijiri and Simon 1977). In Ijiri and Simon's words, these "social phenomena" exhibit "distributions [which] fit quite closely a Pareto distribution." In this chapter we study what others have done in modeling this skew distribution, and we suggest that Simon's models to these social phenomena allow us to explain this usage concentration.

1.1 Empirical Phenomena of Information Usage

1.1.1 80/20 Rule (Pareto Principle)

The simplest way to describe the pattern of usage concentration is to assign some kind of quantitative measurement to it. Vilfredo Pareto (1909) first reported that in Italy about 80% of wealth was concentrated in about 20% of its population. Since then, many other sociological, economic, political, and natural phenomena have been observed to

follow the similar pattern. J. M. Juran, a well-known figure in quality management, claims credit for coining the term Pareto Principle, which is in effect the 80/20 Rule (Sanders 1987). According to Zunde (1984):

> It has been observed that many other empirical phenomena, both in the domain of information science and in other fields, obey this [Pareto] probability distribution law or exhibit dependencies derivable from it.

The 80/20 measure is not a ratio. It has been used mostly as a heuristic to differentiate the "significant few" from the "trivial many," and it was not originally intended as a rule for action (Sanders 1987). For example, approximately 80% of the information usages might involve only about 20% of the resources. Similarly, in libraries (Lancaster and Lee 1985) roughly 80% of transactions involve 20% of holdings. The measure may be 85/35 (85% of sales are generated from 35% of accounts, for instance), 88/40 or 95/25, or any of several other pairings, depending on the point we select to analyze. We may choose the unique point where these two numbers add up to be 100 to describe different usage patterns, thus 70/30 or 90/10. In this example, 90/10 has a higher usage concentration.

The applications of this rule often emphasize the "significant few." For example, Boehm (1987) suggested that 80 percent of rework costs in software development typically result from 20 percent of the problems. The implication is that software verification and validation should focus on identifying and eliminating the high-risk problems in a software project.

1.1.2 Lotka's Law

A high degree of skewness exists in the distribution of output among individuals in certain human activity. The concept of "significant few" and "trivial many" is embedded in "success breeds success;" i.e., successes in many fields tend to center around a few persons, resulting in a relatively small number of people dominating the breakthrough activities in an entire field.

Lotka's law of scientific productivity is an example. In the academic world a frequently cited paper is more likely to be cited again, and a prolific author is more likely to publish again than ones that have published little. In his 1926 paper Lotka examined patterns of scientific productivity among chemists and physicists. He discovered that if he classified this population of scientists according to their publication productivity, then the number of chemists who published n papers was approximately a/n^2, for some positive constant a, i.e.,

$$f(n) = an^{-2}, \quad n = 1,2,3,....$$

Based on this observation, Lotka concluded that the number of persons making 2 contributions is about one-fourth of those making one, the number making 3 contributions is about one-ninth the number making one, and the number making n contributions is about $1/n^2$ of those making one. The proportion of all contributors that make a single contribution is about 60 percent (Lotka 1926). Similar ratios were found in finance and accounting publications (Chung and Cox 1990; Chung, Pak and Cox 1992).

Recently Lotka's law was applied to managing technical innovations. Coile (1988) concluded that an environment that nurtures those rare innovators and encourages somewhat "undisciplined creativity" would be the most beneficial to the company's technical development in the long run, since their successes would most likely bring even more successes. Thus, according to Coile (1988), we should design reward systems that try to support their innovativeness, not stop it.

1.1.3 Bradford's Law

Bradford's law states that a large number of articles related to the same topic seems to concentrate in a few journals. According to Bradford (1934), if a comprehensive literature search is conducted on a subject covering a specified period of time, we often find that the literature is scattered in a regular pattern over a very large number of sources. Further, if we arrange these sources in descending order of productivity (i.e., the journal yielding the most articles at the top of the list and the journals yielding the fewest at the bottom), the sources can be divided into several groups of journals with each group containing the same number of articles. In these succeeding groups, the number of journals will be $1 : j : j^2 : \ldots$, for some constant j. That is, a linear increase in the number of articles requires a geometric increase in the number of journals.

Bradford's law plays a significant role in effective management of library information systems (White and McCain 1989), especially in the area of information retrieval. For example, in systems of limited size (such as ABI/INFORM), the objective

is to include those few important journals which contain a high proportion of the essential articles (Tague 1988).

1.1.4 Zipf's Law

In his 1949 book *Human Behavior and the Principle of Least Effort*, Zipf stated that "if one takes the words making up an extended body of text and ranks them by frequency of occurrence, then the rank r multiplied by its frequency of occurrence, g(r), will be approximately constant." In symbolic form,

$$g(r) = br^{-1}, \qquad r = 1,2,3,...$$

where b is a positive constant whose value depends on the type of text (Zipf 1949). This is usually known as Zipf's first law.

The application of Zipf's first law in IS is frequent. In special-purpose codes such as the family of Huffman codes, a more frequently used character is represented by a shorter bits. Thus structured, Huffman coding can best be used to compress files; however, its variable lengths require time-consuming bit-by-bit examination to decode (Loomis 1989). On the other hand, a fixed-length code would be more convenient to decode, but would have the disadvantage of requiring more storage spaces. Based on Zipf's first law, Thiel and Heaps (1972) designed a scheme to achieve a data compression ratio close to that of the Huffman codes while allowing more rapid decoding of the stored data.

Zipf's first law focuses mainly on words of high frequency (Chen 1989a). In contrast, the formulation of other Zipf's law (often called Zipf's second law) associates with words that rarely occurred (Chen and Leimkuhler 1990, Chen 1989b). Letting f(n)

be the number of words appearing n times in a literary text, then the ratio of f(1) (the number of words occurring once) to the total number of different words in the same text is approximately 0.50. In addition, we have $f(2)/f(1) \approx 0.33$, $f(3)/f(1) \approx 0.17$, $f(4)/f(1) \approx 0.10$, and $f(5)/f(1) \approx 0.07$. In this dissertation we will refer to Zipf's first law simply as Zipf's law, and we will make specific reference to Zipf's second law when necessary.

1.1.5 Observation-Class Relationships

These four empirical findings all show *observation-class* relationship (Chen and Leimkuhler 1986). To be more specific, each empirical finding studies a particular data arrangement. The 80/20 rule in Pareto's form studies the cumulative fraction of observations (e.g., income) and the cumulative fraction of class (e.g., people); Lotka's law relates the observation (e.g., the papers) and the class (e.g., an author) by a frequency-size approach. Bradford's law relates the observation (e.g., the papers) and the class (e.g., a journal) by a cumulative-frequency-log-rank approach. Zipf's law relates the observation (e.g., the word occurrences) and the class (e.g., a word) by a frequency-rank approach.

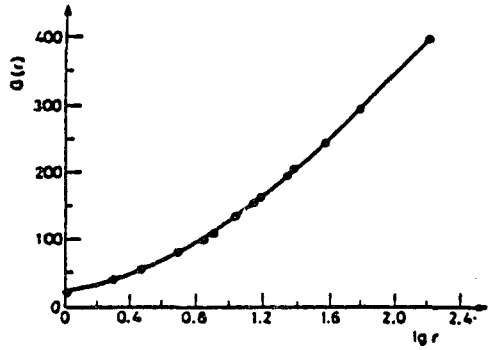
More important, it turned out that underlying all these laws is the same mathematical model — the Pareto distribution (Zunde 1984). Chen and Leimkuhler (1986), through their index approach, proved the mathematical equivalence of Lotka's law, Bradford's law, and Zipf's law. However, there has not been an analytical model for the 80/20 rule. This index approach will be described fully in Chapter 2 to pave the way for our analytical model of the 80/20 rule in Chapter 3.

1.2 Problems of Applying Empirical Findings in IS

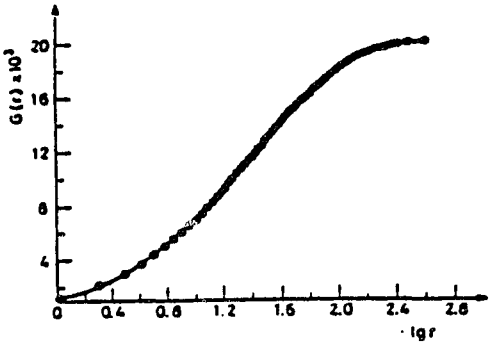
According to Zunde (1984), we call a proposition of science an "empirical law" if "it contains only constructs that refer to observables or are operationally definable ... and laws been extensively verified and found to hold under a variety of conditions." Since these empirical findings hold *in general* and the assessment is only a snapshot of a usually dynamic system, there is no assurance that even if they hold true at a particular time that they would hold at a later date. Use the 80/20 rule for instance. According to a survey of software development managers (Nash 1992), only 5 of the 10 most important systems development issues in 1990 remained in the list in 1992 and with their orders reshuffled. On the other hand, issues that ranked 19 and 17 in 1990 now ranked 10 and 9, respectively, in 1992. Some issues did not even exist in 1990.

These findings may have many exceptions and different forms. For example, Lotka's law holds when the number of journals under study is large; and when broken down to subclasses of journals, the distribution varies according to the quality of journals (Chung, Pak and Cox 1992). In Zipf-type curves, English words have a linear curve, and Latin words form a curve with the concavity to the origin (Chen and Leimkuhler 1987a). Bradford's curve, depending on the data used, may take any of the six general shapes as shown in Figure 1.1; each requires a different formulation (Chen and Leimkuhler 1987b).

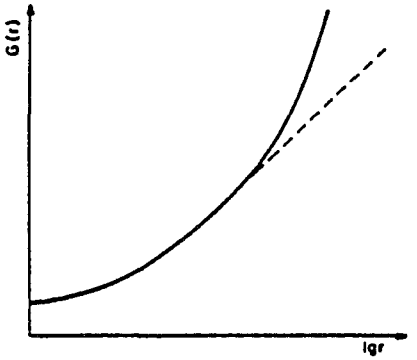
In spite of the problems of stability over time and variations over data, many theories have been developed assuming these findings to be true. Thus, we need to identify factors which influence these findings and the conditions under which these



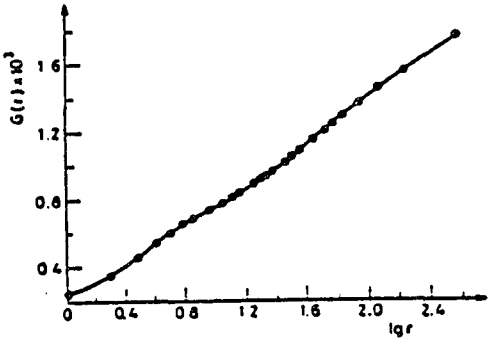
The first class of Bradford-type curves



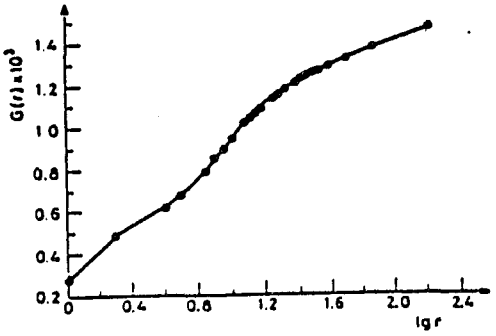
The second class of Bradford-type curve



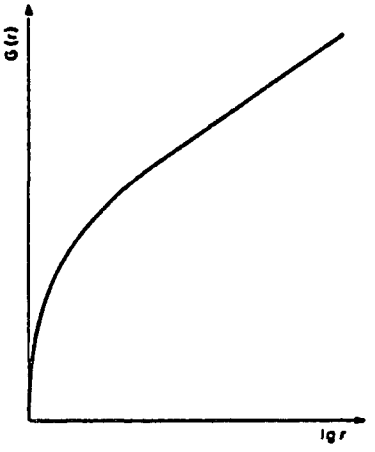
The third class of Bradford-type curves



The fourth class of Bradford-type curves



The fifth class of Bradford-type curves



The sixth class of Bradford-type curves

Figure 1.1: Six Different Classes of Bradford's Curve (Chen and Leimkuhler 1987b)

findings would hold as formulated in order to establish the proper environment in which these findings are applicable.

1.3 Simon's Approach to Empirical Modeling

To identify factors influencing empirical phenomena is a typical example of extreme hypotheses. According to Simon (1968), extreme hypotheses are "assertions that a particular specific functional relationship holds between the independent and the dependent variable." A standard practice for testing extreme hypotheses is the use of goodness-of-fit tests; however, Simon argued that those testing procedures are fundamentally unsatisfactory. He states:

> An extreme hypothesis cannot be sensibly identified with the null hypothesis without shifting completely the burden of proof that is supposed to be assured by a new theory, and what is worse, without making the tacit assumption that the correctness of a theory is an all-or-none matter and not simply a matter of goodness of approximation.

Thus, Simon emphasized that science should be the *discovery* of theories rather than the *testing* of them; therefore, he suggested that theories should rise inductively from data instead of data being collected to fit pre-existing theories.

1.3.1 Empirical Modelling through Successive Refinements

Instead of testing hypotheses, Simon recommended changing the question to that of *estimation*. Ijiri and Simon (1977) stated:

> We are interested in knowing what part of the variance of the data is explained by the theory, and how the remaining variance can be accounted for by successive approximations, rather than in testing whether the variance can be proved to be statistically significant.

Thus, he suggested an approach that consisted of combinations of generalization and refinements. Before one can find phenomena that fit empirical data, one must have

appropriate data that look as though a smooth mathematical function could generate them. "It's the recipe for rabbit stew all over again: First catch the rabbit" (Simon 1991, pp. 372-374).

Simon's (1955) five-step modeling process is listed below:

(1) Begin with empirical data, not hypotheses;

(2) Draw a simple generalization that approximately summarizes striking features of the data;

(3) Find limiting conditions under which deviations from a generalization are small;

(4) Construct simple mechanisms to explain the simple generalizations; and

(5) Propose the explanatory theories that go beyond simple generalizations and create experiments for empirical observations.

Simon's three models of skew distribution are described below. Since his first and second models are quite similar, we refer to them as Simon's Basic Models throughout this dissertation.

1.3.2 Simon's Two Basic Models

Based on Simon's theory of modeling discussed in the previous section, Simon and Van Wormer (1963) proposed a model, in the form of a generating mechanism, which aptly approximated these skew distributions. In terms of simulating information usage patterns, this model would have the following assumptions:

Assumption I: The probability that the $(k+1)$-st information accessed will be an information that has not previously been accessed is $\alpha(k)$, and

Assumption II: The probability that the (k+1)-st information accessed will be an information that has already been accessed i times ($i \geq 1$) is proportional to $i \cdot f(i,k)$, where f(i,k) is the number of distinct information that have been accessed exactly i times each in the first k accesses.

The first assumption differentiates two classes of selections: the information items that have not been selected before ("new" items) and those that have already been chosen ("used" items). The parameter α, therefore, determines whether an item needs to be moved from the class of "no usage" to the class of "used once." If the selection is deemed to be a "used" one, we determine its previous usage through the second assumption. This second assumption essentially describes the property of "success breeds success" by assigning proportionally higher probability of usage to more frequently used information. Thus, a selection that has been used more frequently before is more likely to be selected again.

Simon and Van Wormer (1963) began by assuming the parameter $\alpha(k)$ to be a constant, thus independent of the number of selections, k, that have taken place. They noted that this basic model is only a simple generalization that approximately summarizes striking features of the data. This model of first approximation was further refined by modifying $\alpha(k)$ to be a decreasing function of k: i.e., there is a decreasing probability function $\alpha(k)$, $0 \leq \alpha(k) \leq 1$, that the (k+1)-st selection is chosen for the first time. The values of α in either model are determined by the environment in which the selection process takes place; some environments espouse new selections while others

discourage them. For example, depending on the library and its patrons, rarely used materials may remain just that — rarely used. However, some libraries may have patrons (e.g., scholarly researchers) who actively look for new materials. Their familiarity with library materials and activities of seeking out new information will be reflected in a higher α for this library. In payroll processing, each subsequent selection involves a brand new record, thus we have an extreme case of $\alpha = 1$.

Simon's two basic models use a weak assumption which concerns only with a *class* of information that is used a particular number of times (Ijiri and Simon 1977). It is not required to know *which* information is used how many times. Simon's next refinement, the autoregressive model, does allow the tracking of the usage pattern of individual items.

1.3.3 Simon's Autoregressive Model

The probability of usage for those already-chosen information is assumed to be proportional to its previous usage in Simon's basic models. However, the information not used has a tendency of being "forgotten." In library science it is called "aging" or "decay" (Anderson 1990; Burrell 1980). In other words, the probability of the usage will decrease with time if the book is not in use. Some highly-used library books, once out of fashion, can be neglected for years to come. For analyzing firm sizes which have the same skew distribution as the information usage, Ijiri and Simon (1977) refined the Assumption II of his basic models further to better reflect the reality. This model takes into consideration the recent usage when assigning probability of selection for an "used" information. Simon's autoregressive model is described below.

In terms of information usage, the identity of each information item is maintained from one time period to the next. The selection of the next item is governed by a stochastic process, which depends on how much the item has been used before, and also upon *when* it was last selected. For simplicity in his computations, Simon assumed that only one item is selected at each time period. The probability that an item is chosen next is assumed proportional to a weighted sum of its usage history, where the weight of a usage decreases geometrically, at a rate γ, from the time it was last used. Since the time interval is functionally equivalent to the number of selections, we only indicate the number of selections in lieu of time interval in this dissertation.

Simon formalized these notions as follows: Let $y_j(k)$, equals to 1 or 0, be our selection of j-th information item during the k-th selection, then the total usage of the j-th item at the end of the k-th selection is simply $\sum_{\tau=1}^{k} y_j(\tau)$. If we denote T to be the total number of distinct items that have been selected at least once at the end of the k-th selection, then the expected usage of the j-th item for the (k+1)-st selection is

$$p[y_j(k+1) = 1] = \frac{1}{W_k} \sum_{\tau=1}^{k} y_j(\tau)\gamma^{k-\tau}$$

where

$$W_k = \sum_{j=1}^{T} \sum_{\tau=1}^{k} y_j(\tau)\gamma^{k-\tau}.$$

W_k is the same for all items.

1.4 Dissertation Contributions and Organization

The contributions of this dissertation is threefold. First, we conduct analytical examination of the 80/20 rule in order to find the characteristics of the distribution of

information usage. Second, we simulate information usage patterns using Simon's generating mechanisms to see the effect of different factors on the 80/20 rule and other empirical phenomena. We find that different forms of empirical phenomena can be reproduced by different combinations of parameter values. Based on results of our analytical model for the 80/20 rule, we also estimate the parameter values of empirical data and then use these values to simulate future usage patterns. Third, we use the weeding policy in the library resource management as a case study to demonstrate the applicability of Simon's usage model in improving the productivity of information systems.

The organization of this dissertation is as follows. Chapter 2 is the presentation of the index approach and Simon's generating mechanism. In this chapter we first define the notations used in this dissertation, and we explain Chen's index approach which is essential in establishing the relationships between the 80/20 rule and Simon's model. We also review the indexed formulations of Lotka's law, Bradford's law, and Zipf's laws — which have already been conducted in the literature. Thereafter, we describe the simulation models and algorithms used in examining the important parameters of these empirical phenomena.

The contributions of this dissertation are presented from Chapter 3 on. Chapter 3 is the analytical analysis which, based on the index approach, provides a theoretical foundation for the 80/20 rule. The resultant formulation requires less assumptions and parameters. Chapters 4 and 5 are in-depth discussions of applications of Simon's model to the simulation of the four empirical phenomena described in this chapter.

Specifically, Chapter 4 examines Simon's two basic models and Chapter 5 the autoregressive model. Here we discuss the simulation results of altering the three parameters (the probability of new entry, α; the decay factor, γ; and the total number of iteration, N), and how different versions of empirical phenomena can be simulated through different values of these parameters. The significance of the findings in these two chapters is two fold: first, we demonstrate that Simon's model provides a unifying theoretical foundation for all of these empirical phenomena in question; and second, we now have a means to estimate the values of parameters through observing the usage pattern of a particular set of data. Furthermore, Chapters 4 and 5 provide limiting conditions in which these empirical phenomena hold. Chapter 6 uses actual usage pattern of a regional university to test our findings, and we demonstrate that Simon's model can provide theoretical support to the weeding policy used in library resource management. Finally, Chapter 7 is the conclusion.

CHAPTER 2

LITERATURE REVIEW ON THE INDEX APPROACH AND SIMON'S GENERATING MECHANISM

As part of the literature review, in this chapter we first describe the index approach proposed by Chen and Leimkuhler (1986) and the notations used; then we discuss Simon's generating mechanism as the backbone of simulation in studying empirical phenomena. These two tools are essential in the development of our analytical model and simulations in this dissertation.

2.1 The Index Approach

2.1.1 Notations and Examples

Let us take Kendall's (1960) study on 1763 papers published on operational research (Table 2.1, columns denoted n_i and $f(n_i)$) as an example to understand these empirical phenomena. If we tabulate the number of authors who have published n papers and arrange this list in the ascending order of n, we would find that n does not run consecutively at places, especially when n is large. We would also find that there are m different clusters of authors who publish the same number of papers, and $m \leq \max\{n\}$. To take into account the scatter of the larger values of n, Chen and Leimkuhler (1986) introduced an index $i = 1,2,\ldots,m$, for the m successive observations of n and let n_i denote the i-th nonzero value of n where $n_i < n_{i+1}$.

Table 2.1: Demonstration of Empirical Laws Using Kendall's Data (1960)

i	n_i	$f(n_i)$	$n_i f(n_i)$	r_i	$G(r_i)$	x_i	θ_i	$\log(r_i)$	$\log(g(r_i))$
1	1	203	203	1	242	0.003	0.137	0.000	2.384
2	2	54	108	2	356	0.005	0.202	0.301	2.057
3	3	29	87	3	458	0.008	0.260	0.477	2.009
4	4	17	68	4	553	0.011	0.314	0.602	1.978
5	5	10	50	5	611	0.014	0.347	0.699	1.763
6	6	6	36	6	660	0.016	0.374	0.778	1.690
7	7	8	56	7	694	0.019	0.394	0.845	1.531
8	8	8	64	9	738	0.024	0.419	0.954	1.342
9	9	4	36	11	780	0.030	0.442	1.041	1.322
10	10	3	30	13	820	0.035	0.465	1.114	1.301
11	11	5	55	14	838	0.038	0.475	1.146	1.255
12	12	2	24	18	902	0.049	0.512	1.255	1.204
13	14	1	14	20	932	0.054	0.529	1.301	1.176
14	15	2	30	21	946	0.057	0.537	1.322	1.146
15	16	4	64	23	970	0.062	0.550	1.362	1.079
16	18	1	18	28	1025	0.076	0.581	1.447	1.041
17	20	2	40	31	1055	0.084	0.598	1.491	1.000
18	21	2	42	35	1091	0.095	0.619	1.544	0.954
19	22	2	44	43	1155	0.116	0.655	1.633	0.903
20	34	1	34	51	1211	0.138	0.687	1.708	0.845
21	49	1	49	57	1247	0.154	0.707	1.756	0.778
22	58	1	58	67	1297	0.181	0.736	1.826	0.699
23	95	1	95	84	1365	0.227	0.774	1.924	0.602
24	102	1	102	113	1452	0.305	0.824	2.053	0.477
25	114	1	114	167	1560	0.451	0.885	2.223	0.301
26	242	1	242	370	1763	1.000	1.000	2.568	0.000
		---	-----						
m=26		T=370	N=1763	μ=4.7649					

80/20: θ_i vs. x_i; % cumulative transactions vs. % cumulative holdings; x_i is r_i/T and θ_i is $G(r_i)/N$.

Lotka: $f(n_i)$ vs. n_i; number of publication vs. number of authors.

Bradford: $G(r_i)$ vs. $\log(r_i)$; cumulative papers at rank vs. log of rank.

Zipf: $\log(g(r_i))$ vs. $\log(r_i)$; log of word frequency vs. log of the rank; $g(r_i)$ is essentially n_{m-i+1}.

In terms of information usage, we define:

m = the maximum number of clusters of items with the same usage;

n_i = the number of times an information item is used, $i = 1,2,...m$;

$f(n_i)$ = the number of items that have been used n_i times;

$F(n_i)$ = the number of items that have been used no less than n_i times;

$r_i = \sum_{k=m-i+1}^{m} f(n_k)$ = the rank of item i; ranked according to its usage;

$g(r_i) = n_{m-i+1}$ = the number of times an item with rank r_i was used;

$G(r_i) = \sum_{k=m-i+1}^{m} n_k f(n_k)$ = the total number of usages for items ranking no greater than r_i;

$T = \sum_{i=1}^{m} f(n_i)$ = the total number of different items;

$N = \sum_{i=1}^{m} n_i f(n_i)$ = the total number of usages; and

$x_i = r_i/T$ = the fraction of total items which have been used at least n_i times;

$\theta_i = G(r_i)/N$ = the fraction of total usage for the first i items;

$\mu = N/T$ = the average usage per item.

Table 2.1 uses Kendall's (1960) data to demonstrate how parameters used in these empirical phenomena are transformed from the raw data. This table is the basis of the following graphs: the 80/20 rule (Figure 2.1a), Lotka's law (Figure 2.1b), Bradford's law (Figure 2.1c), and Zipf's law (Figure 2.1d).

The column x_i is the percent of total holding at this rank. For example, ranks 1 through 67 (meaning that there are 67 authors who published at least 5 papers) is 18.1% of the total 370 authors examined. The column θ_i shows the cumulative activities as a percent of the total usage. For example, by rank 67 there have been 1297 papers

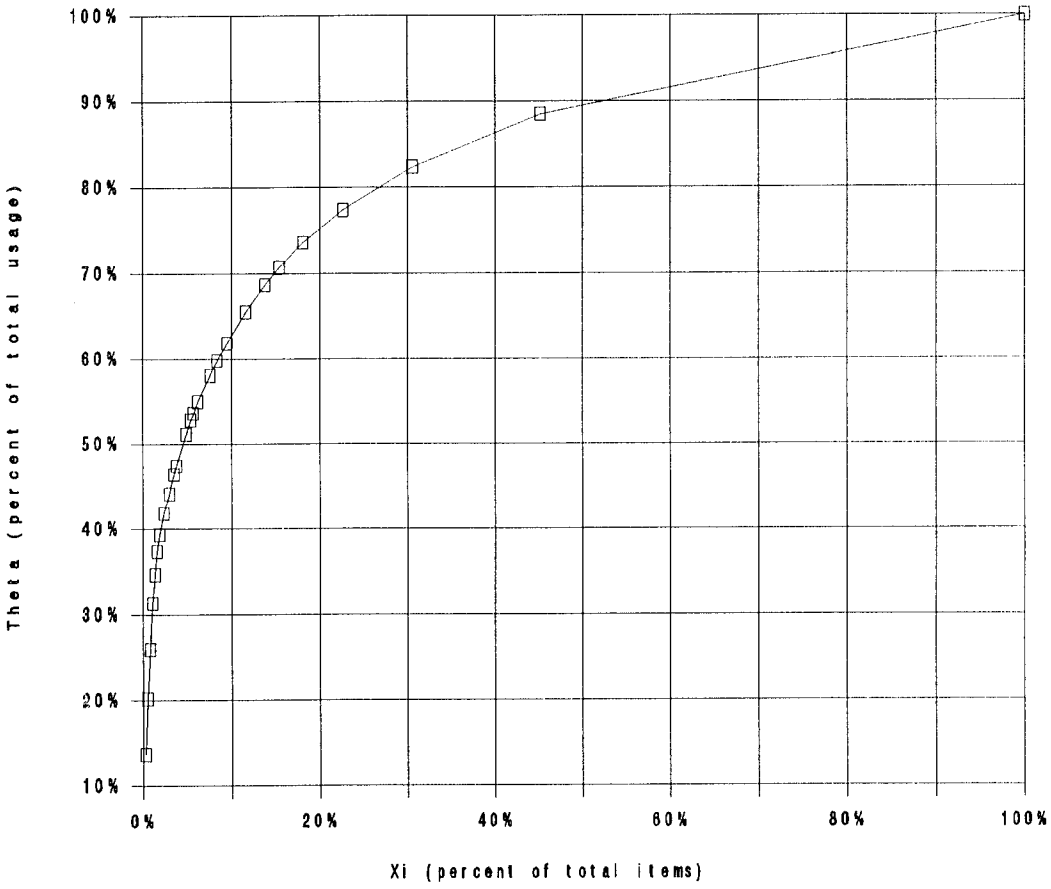


Figure 2.1a: 80/20 Rule Using Kendall's Data

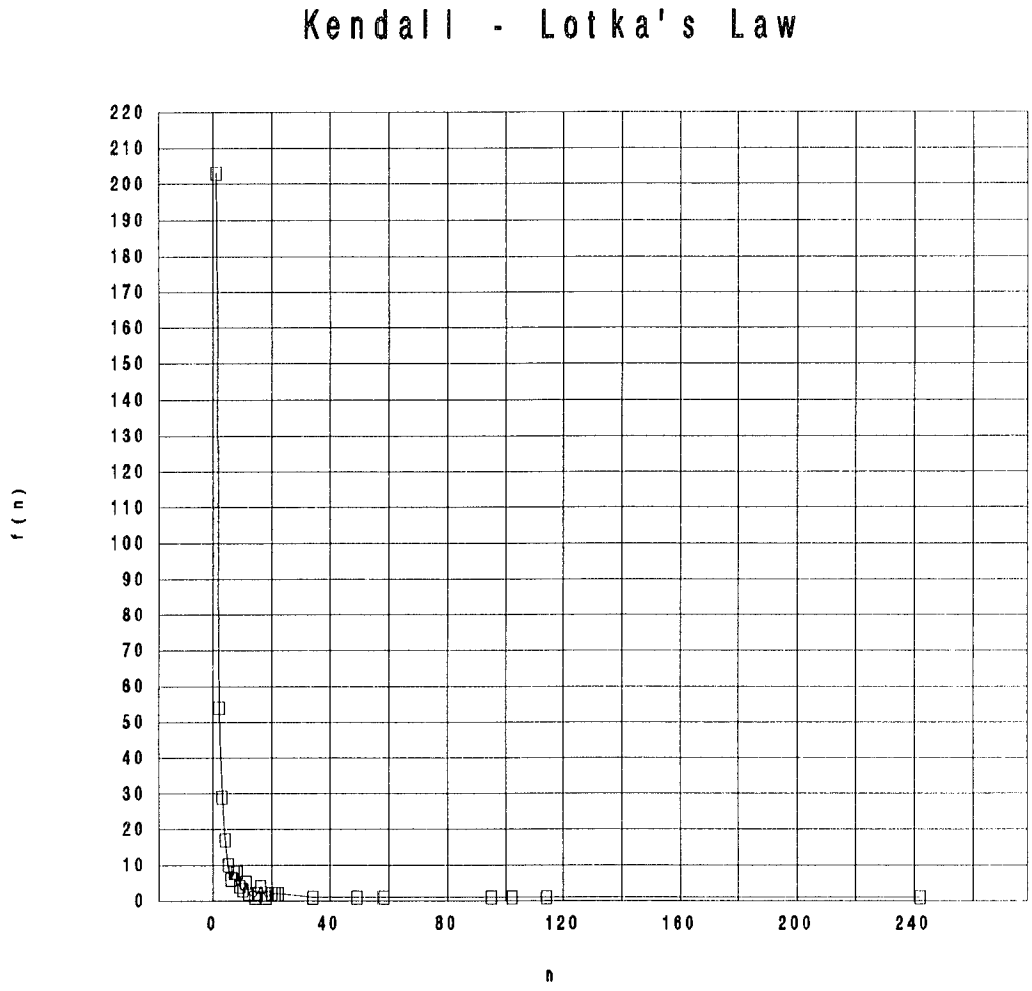


Figure 2.1b: Lotka's Law Using Kendall's Data

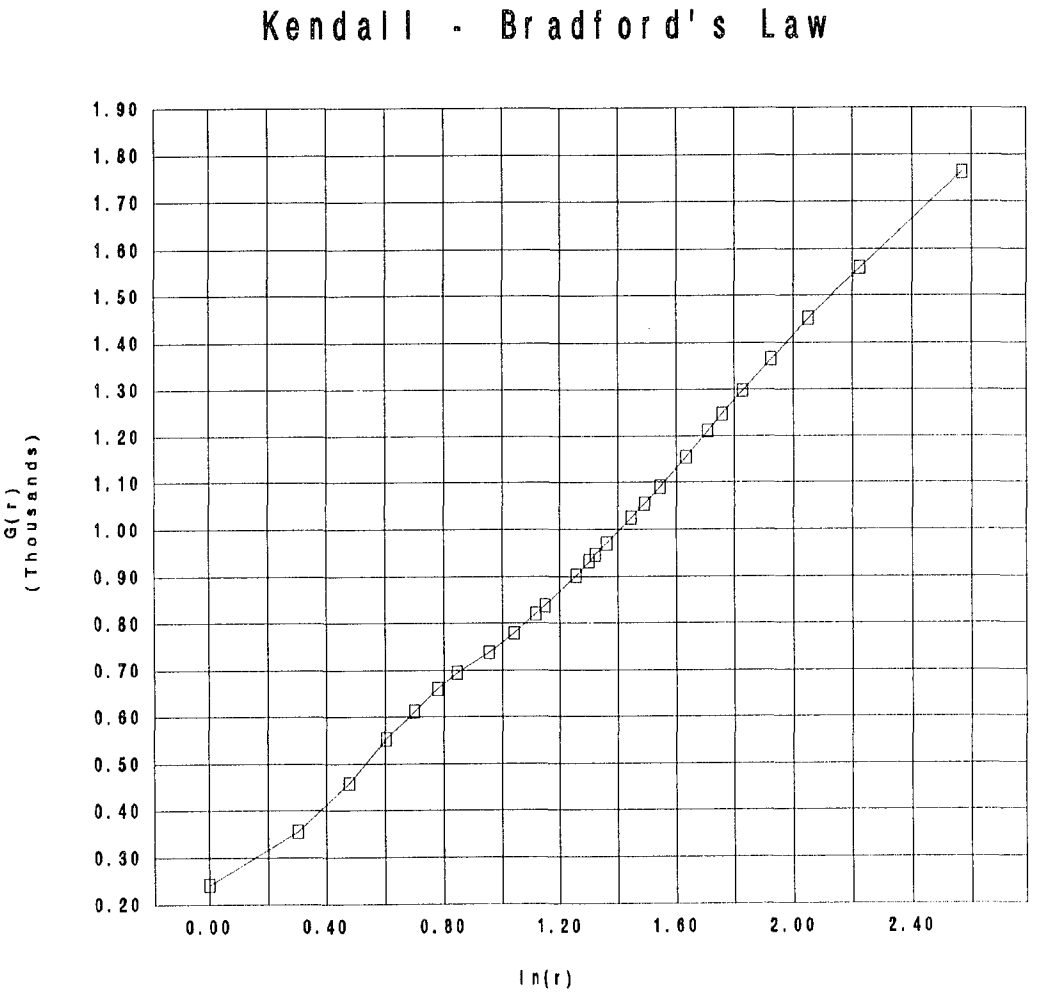


Figure 2.1c: Bradford's Law Using Kendall's Data

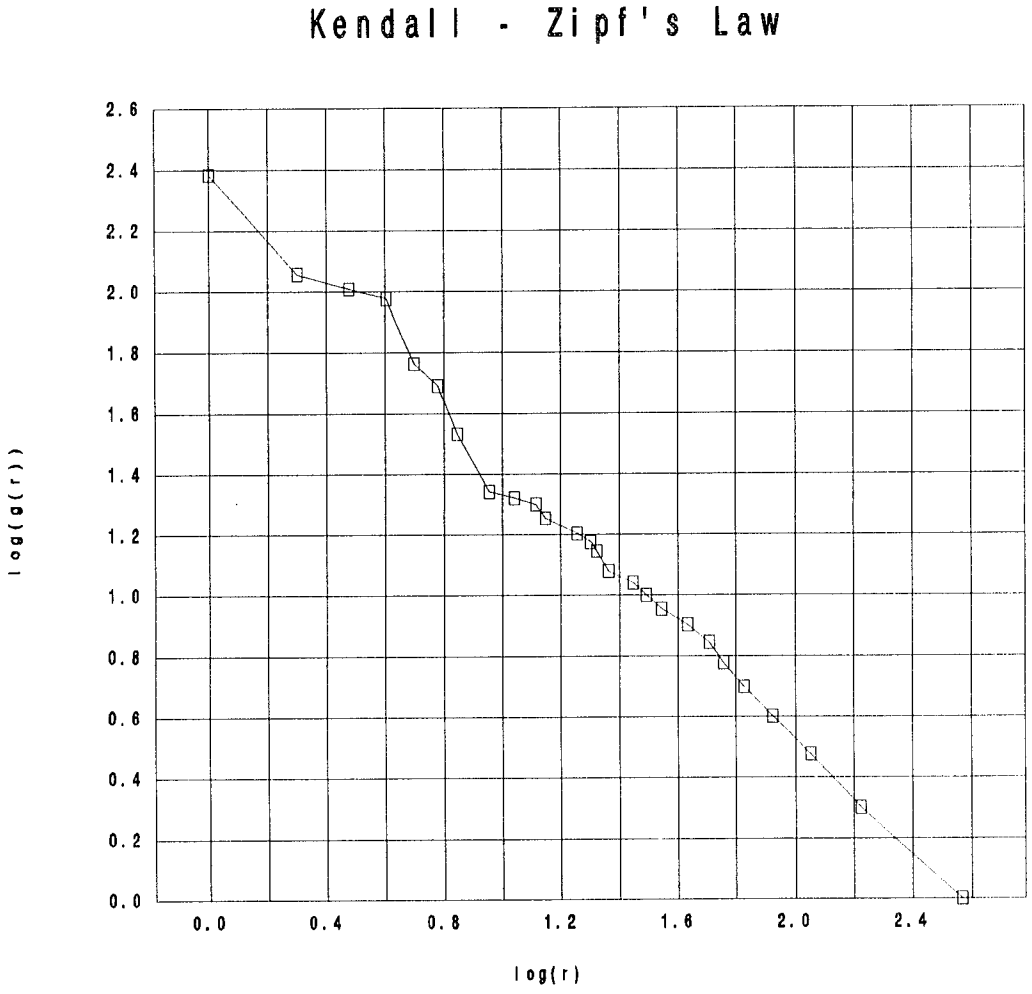


Figure 2.1d: Zipf's Law Using Kendall's Data

published (67 authors published 1297 papers), and that is 73.6% of the total 1763 papers studied. In this particular case, the concentration measurement is 74/18 (74% of papers are published by 18% of authors), or we may examine Figure 2.1a and describe the curve to be 78/22, which is rather close to 80/20.

Lotka's curve is drawn using $f(n_i)$ vs. n_i (the number of authors vs. the number of paper each of them publishes). Suppose we use a modified form of this curve which uses $\log(f(n_i))$ and $\log(n_i)$ instead; since $f(n_i) = an_i^{-2}$, we expect the shape of the curve to be linear with a slope of -2. For Bradford type of curves, we have $G(r_i)$ vs. $\log(r_i)$; therefore, for a linear growth of $G(r_i)$ (total number of usage) and a geometric growth of r_i (the number of items required), the curve is expected to be linear with a positive slope. Zipf's curves are plotted using $\log(g(r_i))$ (log of the number of usage for the item) vs. $\log(r_i)$ (the log of the item's rank). Since $g(r_i) = br_i^{-1}$, we expect to have the curve to be linear with a slope of -1.

Notice that the skew distribution has two extremes — only a few items that are used many times, and many items that are used only a few times — and those in between. Each of these empirical phenomena has concerns with different parts of the same set of usage data. In the 80/20 rule, we are more interested in those few items that are frequently used, while Lotka's law places emphasis on the large number of authors who publish only a few papers. Bradford's law shows the connection between these two extremes in terms of clustering the number of papers. Zipf's first law is mainly concerned with the relationship between high-usage items and their ranks, and Zipf's second law shows the ratio relationship between those many items that are used only a

few times. However, all these observations are based on different transformation of the same set of usage data.

2.1.2 Three Significant Clusters

The index approach (Chen and Leimkuhler 1986) identifies three significant clusters of n: (a) where i is small and $n_i = i$; (b) where i is large and $f(n_i) = 1$; and (c) those in between. Let i_ℓ be the largest i where $n_i = i$, and let i_u be the smallest i where $f(n_i) = 1$, $i = u, u+1, \ldots, m$, and $f(n_{u-1}) \neq 1$ to define the following properties:

$$\begin{aligned} &\text{cluster 1:} && n_i = i,\ 1 \le i \le i_\ell, \\ &\text{cluster 2:} && n_i \approx i \text{ and } f(n_i) \approx 1,\ i_\ell + 1 \le i \le i_u - 1, \\ &\text{cluster 3:} && f(n_i) = 1,\ i_u \le i \le m. \end{aligned} \tag{2.1}$$

For example, $n_i = i$ for $1 \le i \le 10$ (cluster 1); $n_i \approx i$ and $f(n_i) \approx 1$ for $i = 11$ (cluster 2); and $f(n_i) = 1$ for $12 \le i \le 14$ (cluster 3) in Table 2.1(a). In Table 2.1(b) clusters 1, 2, and 3 are $1 \le i \le 12$; $13 \le i \le 19$; and $20 \le i \le 26$, respectively. Notice that for small i, $n_i = i$; and for $i \approx m$, $f(n_i) = 1$. Correspondingly cluster 1 contains the items of low usage while cluster 3 items have the highest usage frequency. These three clusters are essential in the analytical examination of the empirical phenomena.

2.2 Applying Index Approach to Empirical Laws

This section summarizes findings of Chen and Leimkuhler (Chen and Leimkuhler 1986, 1987a, 1987b; Chen 1989) on Lotka's law, Bradford's law, and Zipf's laws.

2.2.1 Mathematical Equivalence of Empirical Laws

Chen and Leimkuhler (1986) used the index approach and showed that the three empirical laws are mathematically equivalent. That is, for i = 1,2,..,m, Equations 2.2, 2.3, and 2.4 have the following relationship:

$$F(n_i) = dn_i^e - b \tag{2.2}$$

$$\text{iff} \quad G(r_i) = a \sum_{k=1}^{i} [r_k^c (r_k - r_{k-1})] \tag{2.3}$$

and

$$\text{iff} \quad g(r_i) = a(r_i + b)^c, \tag{2.4}$$

where a, b, c, d, e are constants and a,d > 0; c,e < 0; ce = 1, $da^e = 1$, and b > -1. These three equations, without the index notations, are general formulations of Lotka's law (Chen 1989), Bradford's law (Chen and Leimkuhler 1987b), and Zipf's law (Chen and Leimkuhler 1987a), respectively. Equations 2.2 and 2.4 will be used later to help derive the indexed version of the 80/20 rule in Chapter 3.

This result not only shows that these three laws are basically different ways of looking at the same phenomenon but also enlarges the number of tests that might be applied to a particular set of data (Chen and Leimkuhler 1986). Furthermore, since Lotka's law deals with the most original form of data (n_i and $f(n_i)$), the effort of modeling the following three empirical phenomena can be reduced to modeling Lotka's law. This will be addressed in Section 2.2.4.

2.2.2 Bradford's Law

Through the index approach, Chen and Leimkuhler (1987b) studied the three clusters as depicted in Equation 2.1. They concluded that the six different classes of

Bradford as illustrated in Figure 1.1 can be explained by the three corollaries in their study. Assuming the index version of Lotka's law holds, they found that cluster 2 is approximately linear. The concavity, linearity, or convexity of cluster 1 and 2 are determined by h_i and e, respectively, where

$$h_i = \frac{n_{m-i}}{n_{m-i-1}} - \frac{\log \frac{i+1}{i}}{\log \frac{i+2}{i+1}},$$

and e is defined as in the previous section.

2.2.3 Zipf's Two Laws

Although Zipf's curves all show a linear decreasing pattern to the right of the graph, a segment of the curve can be either convex or concave to the origin. Through analyzing the slopes of Zipfian curves using Equation 2.4 (indexed version of Zipf's law), Chen and Leimkuhler (1987a) show that the Zipf-type curve can be concavely decreasing, linearly decreasing, or convexly decreasing, depending on the value of b. As in Bradford's law, cluster 2 is nearly linear. Cluster 3 is also linear and has a lope equal to c. The indexed version of Zipf's law takes explicit account of the sequence of observed values of the variables and makes it possible to account for the variations normally encountered with Zipf-type data (Chen and Leimkuhler 1987a).

Using definition of r_i and $g(r_i)$, Chen and Leimkuhler (1990) derived $F(n_i) = d - b$, and $f(n_i) = d(i^c - (i+1)^c)$, $i \ll m$; which implies the Zipf's second law when $c = -1$, $b = 0$, and $i = 1,2,3,4,5$.

2.2.4 Lotka's Law

Since Bradford's law, Zipf's law, and Lotka's law are mathematically equivalent, the study comes down to the stochastic modeling of Lotka's law because of its simplicity. Chen (1989) adapts Simon's five-step modeling process to study Lotka's law. Using the index approach, Chen finds that in cluster 1, where $n_i = i$,

$$f(i) = d(i^c - (i+1)^c);$$

and in cluster 3 where $f(n_i) = 1$,

$$n_i = a(m - i + 1 + b)^c.$$

Chen (1989) shows that the empirical phenomena are marginal properties of the time series he has studied, and the two influential variables are the entrance of new journals and the productivity of old journals. Since Simon and Van Wormer's (1963) generating mechanism (Simon's model) incorporates the concept of new and old entities, Chen (1989) further shows that Equation 2.2 can be derived from the expected value derived from the two assumptions in Simon's basic models. Thus, Simon's generating mechanism provides a theoretical foundation for these empirical phenomena.

2.3 Simon's Generating Mechanism

2.3.1 The Algorithm for Simon's Two Basic Models

The two versions of Simon's model can be easily programmed on a computer to simulate empirical data $(n_i, f(n_i))$, $i=1,2,\ldots,m$. The simulation algorithm consists of two steps (Simon and Van Wormer 1963):

Step 1: For each selection k ($1 \leq k \leq N$), we generate a random number a from the uniform distribution with range 0 to 1. If $a \leq \alpha(k)$,

$f(1,k) = f(1,k-1) + 1$ — i.e., this is a "new" item and we add it to the "used once" category; otherwise we go to step 2 to find out its usage history.

Step 2: A random number b is drawn from the uniform distribution with range $1 \leq b \leq k$. Begin with $j = 1$, the cumulant of $j \cdot f(j,k-1)$ is computed to find an n such that $\sum_{j=1}^{n} j\, f(j,k-1) \geq b$. Then $f(n,k) = f(n,k-1) - 1$, and $f(n+1,k) = f(n+1,k-1) + 1$. This is equivalent to saying that the k-th item selected was used n times before, and now it is used $n+1$ times.

2.3.2 The Algorithm for Simon's Autoregressive Model

In the autoregressive model, the selection of the k-th item is made in two stages:

Stage I: This stage is similar to the Step 1 stated in the previous section.

Stage II: This stage is a refinement of the Step II in the basic models. The weight of previous usages are changed by a factor γ at each selection. Since the weight of the usage during period (k-1) is 1, this weight will become $\gamma^{(\tau-1)}$ by period $(k-\tau)$. The sum of these weights, i.e., $W(k-1) = \Sigma_j w_j(k-1)$, are kept for all items. We draw a random number b from a uniform distribution between 0 and W(k-1), and assign the k-th selection to the j-th item, where j is the smallest integer which satisfies $\sum_{i=1}^{j} w_i(k-1) \geq b$.

2.3.3 Initial Conditions of the Simulation Models

The simulation data for Simon's basic models were obtained from a computer program written in Turbo Pascal running on a 386 personal computer. To start the simulation program, the initial conditions f(n,0), n = 1,...,N, were provided. Since moderate changes in the initial conditions do not appear to affect the equilibrium distributions (Simon and Van Wormer 1963), we set the initial conditions with f(1,0) = 3, and f(n,0) = 0 for n = 2,...,N; i.e., there already exist three items with one previous usage each.

In addition to these initial conditions, the simulations were carried out with N ranging from 1,000 to 30,000. On the other dimension, α also was varied. For constant α, it ranged from 0.1 to 0.9 with 0.1 increment plus the two extreme conditions of α = 0.01 and α = 0.99. For the decreasing function, we used α = A/ln(R), R = 1,2,...N, where A ranged from 1.00 to 2.00 with 0.25 increment. A sample program is listed in Appendix A. Once the usage pattern is generated, the data (i.e., n_i and $f(n_i)$) are entered into a Lotus program setup similar to that of Table 2.1, from which we generate graphs for the empirical phenomena.

For the autoregressive models, since stage II of the process requires that some "used" items exist, we set the initial conditions the same as those in basic models. The final distribution is not entirely independent of the initial conditions but tends to become independent as N grows large. Because of the amount of memory space required for tracking activities of individual items in the autoregressive model, the simulations were run on an IBM 9370 mainframe computer, using IBM VS Pascal Release 2. The

program is listed in Appendix B. Since this model generates individual usage frequencies, they are sorted and tabulated to generate a summary of $(n_i, f(n_i))$ pairs. This summary data are then processed the same way as those in basic models. For reasons to be discussed in Chapter 5, most of the results of autoregressive model were obtained using $\alpha = 0.20$ and $N = 20{,}000$.

CHAPTER 3

THE ANALYTICAL STUDY OF THE 80/20 RULE

We apply the index approach to analytically examine the 80/20 rule in this chapter.

3.1 On the 80/20 Rule

Burrell (1985) studied the 80/20 rule and found that the minimum holdings needed for 80% of the circulation varies inversely with the average circulation rate and is usually greater than 20%. Using notations as defined in Section 2.1.1 (replacing "holdings" for "information items" and "circulation" for "usage"), by assuming f(n) follows a negative binomial distribution Burrell (1985) derived the relationship:

$$\theta(x,\mu) = x + \frac{x \log x}{\mu \log[(\mu-1)/\mu]} \tag{3.1}$$

Subsequently, Egghe (1986) showed that the minimum holdings is close to 20% when the frequency follows Lotka's law, i.e., $f(n) = a/n^2$, and derived the relationship:

$$\theta(x,\mu) = 1 - (\frac{6}{\pi^2 \mu})\,[E + \log(\frac{6}{\pi^2 x})] \tag{3.2}$$

where $E = 0.57722...$ is Euler's number. Both formulations exhibit the inverse property: given $\theta = 0.80$, x increases if and only if μ decreases.

Both Burrell and Egghe's approaches assume that the values of n run consecutively from 1 to N. However, as we have pointed out in Section 2.1.1, in real data the circulation values observed for the more frequently used items tend to jump

erratically until reaching the largest circulation value. The following analysis uses the index approach as described in Section 2.1 and does not require the above-mentioned assumptions.

3.1.1 General Formulas

Theorem 1: For i = 1,2,...m,

$$\theta_i = x_i \frac{\mu_i}{\mu}. \tag{3.3}$$

Proof: Let x_i be the fraction of holdings, θ_i be the fraction of circulation, and μ_i be average circulation per holding for the top i holdings, so that

$$x_i = \frac{1}{T} \sum_{k=m-i+1}^{m} f(n_k), \tag{3.4}$$

$$\theta_i = \frac{1}{N} \sum_{k=m-i+1}^{m} n_k f(n_k) \tag{3.5}$$

and

$$\mu_i = \frac{\sum_{k=m-i+1}^{m} n_k f(n_k)}{\sum_{k=m-i+1}^{m} f(n_k)} = \frac{\mu\theta_i}{x_i}. \tag{3.6}$$

By rearranging Equation 3.6 we obtain Equation 3.3. [ ]

From this relationship between μ and μ_i, we can immediately conclude that for the 80/20 rule to be true there must exist an i, $1 \leq i \leq m$, such that $(x_i,\theta_i) = (0.2,0.8)$ and $\mu_i = 4\mu$; i.e., the average circulation per holding for the top class of holdings is four times the average circulation for all holdings.

In the following theorem, we use interpolation to derive the exact formulation of the 80/20 curve.

Theorem 2: For i = 1,2,...,m,

$$\theta = x_{j-1}\frac{\mu_{j-1}}{\mu} + (x - x_{j-1})\frac{n_{m-j+1}}{\mu}, \tag{3.7}$$

where x is given and j is the smallest i, such that $x_i \geq x$.

Proof: Let j be the smallest i, such that $x_i \geq x$ and s_i is the slope of the line segment between (x_{j-1},θ_{j-1}) and (x_j,θ_j), i = 1,2,...,m and $(x_0,\theta_0) = (0,0)$. Given x,

$$\theta = \theta_{j-1} + (x - x_{j-1})\, s_j, \tag{3.8}$$

and from Equations 3.4 and 3.5

$$s_j = \frac{\theta_j - \theta_{j-1}}{x_j - x_{j-1}} = \frac{T}{N} n_{m-j+1} = \frac{n_{m-j+1}}{\mu}. \tag{3.9}$$

Equations 3.8 and 3.9 derive Equation 3.7. [ ]

Equation 3.6 is a special case of Equation 3.7 when $x = x_{j-1}$. Equation 3.7 can be rewritten as follows:

$$x = \frac{\mu\theta - x_{j-1}(\mu_{j-1} - n_{m-j+1})}{n_{m-j+1}}, \tag{3.10}$$

where θ is given and j is the smallest i such that $\theta_i \geq \theta$. Equation 3.10 shows that the 80/20 rule is determined by four types of parameters: μ, j, μ_{j-1}, and n_{m-j+1}, for all j, which depend on the value of parameter m, the distribution of $f(n_i)$, and the scattering pattern of n_i, for i = 1,2,...m.

A simpler alternative for analyzing the 80/20 rule is to examine the scattering pattern of slope-distance pairs of the data. Let s_i and d_i be the slope and distance of the line segment between (x_{i-1},θ_{i-1}) and (x_i,θ_i), i = 1,2,...,m, respectively, and $(x_0, \theta_0) = (0,0)$, then we have

Theorem 3: For i = 1,2,...,m,

$$s_i = \frac{n_{m-i+1}}{\mu} \tag{3.11}$$

and

$$d_i = \frac{1}{N}\sqrt{(\mu^2 + n_{m-i+1}^2)\, f^2(n_{m-i+1})} \tag{3.12}$$

Proof: From Equations 3.4 and 3.5, we obtain

$$s_i = \frac{\theta_i - \theta_{i-1}}{x_1 - x_{i-1}} = \frac{T}{N} n_{m-i+1} = \frac{n_{m-i+1}}{\mu},$$

and

$$\begin{aligned} d_i &= \sqrt{(x_i - x_{i-1})^2 + (\theta_i - \theta_{i-1})^2} \\ &= \sqrt{\frac{f^2(n_{m-i+1})}{T^2} + \frac{n_{m-i+1}^2 f^2(n_{m-i+1})}{N^2}} \\ &= \frac{1}{N}\sqrt{(\mu^2 + n_{m-i+1}^2 f^2(n_{m-i+1})} \end{aligned}$$

[ ]

Since (s_i,d_i) uniquely determines (x_i,θ_i) and vice versa, for $i = 1,2,\ldots,m$, and for each i, (s_i,d_i) has much simpler formulation than (x_i,θ_i), we will focus on the set of (s_i,d_i), $i = 1,2,\ldots,m$, in the rest of this section. Theorem 3 and the three properties shown in Equation 2.1 enable us to derive immediately the following corollary:

Corollary 1:

(a) For $1 \le i \le m\text{-}i_u+1$,

$$s_i = \frac{n_{m-i+1}}{\mu}$$

and

(b) For $m\text{-}i_u+2 \leq i \leq m\text{-}i_\ell$,

$$s_i \simeq \frac{n_{m-i+1}}{\mu}$$

and

$$d_i \simeq \frac{1}{N}\sqrt{\mu^2 + n_{m-i+1}^2} \,. \tag{3.14}$$

(c) For $m\text{-}i_\ell+1 \leq i \leq m$,

$$s_i = \frac{m-i+1}{\mu}$$

and

$$d_i = \frac{1}{N}\sqrt{(\mu^2 + (m-i+1)^2)\, f^2(m-i+1)} \,. \qquad [\,] \tag{3.15}$$

Let us define the three categories of s_i and d_i — where $1 \leq i \leq m\text{-}i_u+1$, $m\text{-}i_u+2 \leq i \leq m\text{-}i_\ell$, and $m\text{-}i_\ell+1 \leq i \leq m$ — to be region I, region II, and region III, respectively. Equation 3.13 indicates that the shape of the curve in region I depends on the value of μ, N, and the scattering pattern of n_j, $i_u \leq j \leq m$. Since n_j is an increasing function for $i_u \leq j \leq m$, we see that s_i and d_i to be decreasing for $1 \leq i \leq m\text{-}i_u+1$. Equation 3.14 indicates that the shape of the curve in region II depends on m, μ, and N, which are all constants. As such, s_i and d_i are decreasing for $m\text{-}i_u+2 \leq i \leq m\text{-}i_\ell$. Finally, Equation 3.15 indicates that the values m, μ, N, and f(j), $1 \leq j \leq i_\ell$, determine the shape of the curve, where s_i is a decreasing function with respect to i, for $m\text{-}i_\ell+1 \leq i \leq m$. Since f(j) is a strictly decreasing function for $1 \leq j \leq i_\ell$, we see that d_i shows an increasing pattern for $m\text{-}i_\ell+1 \leq i \leq m$.

3.1.2 Three Significant Regions

Region I: the Significant Few

Equation (3.13) implies that

$$(s_1,d_1) = \left(\frac{n_m}{\mu}, \frac{1}{N}\sqrt{\mu^2+n_m^2}\right),$$
$$(s_2,d_2) = \left(\frac{n_{m-1}}{\mu}, \frac{1}{N}\sqrt{\mu^2+n_{m-1}^2}\right),$$

. . .

and (3.16)

$$(s_{m-i_u+1},d_{m-i_u+1}) = \left(\frac{n_{i_u}}{\mu}, \frac{1}{N}\sqrt{\mu^2+n_{i_u}^2}\right).$$

Since n_j is an increasing function for $i_u \leq j \leq m$, we see that s_i and d_i are decreasing for $1 \leq i \leq m\text{-}i_u+1$. Furthermore, the shape of the curve in region I depends on the value of μ, N, and the scattering pattern of n_j, $i_u \leq j \leq m$. An immediate implication of the slope-distance pairs above is that the 80/20 rule does not hold if $s_1 < 4$. On the other hand, the 80/20 rule holds if $s_1 = 4$ and $d_1 = \sqrt{0.68}$.

Recalling Equation 2.1, the region where $i_u \leq i \leq m$ and $f(n_i) = 1$ contains the elements of the "significant few." One of the most cited laws of the significant few is based on Zipf's rank-frequency approach. By applying Zipf's rank-frequency approach and Equation 2.4, we can simplify the slope-distance pairs in region I as follows:

Corollary 2: If the indexed formulation of Zipf's law holds, then the slope-distance pairs of the curve in region I are:

$$(s_i,d_i) = \left(\frac{a(i+b)^c}{\mu},\ \frac{1}{N}\sqrt{\mu^2 + a^2(i+b)^{2c}}\ \right), \tag{3.17}$$

$i = 1,2,\ldots,m\text{-}i_u+1.$

Proof: Since $n_m > n_{m-1} > \ldots > n_{i_t}$, the ranks correspond to n_m, m_{m-1}, ..., n_{i_t}, are $1,2,\ldots, m\text{-}i_u+1$, respectively. If the indexed formulation of Zipf's law holds, then

Rank	Frequency based on Zipf's Law	
1	n_m	$= a(1+b)^c$
2	n_{m-1}	$= a(2+b)^c$
.	.	
.	.	
.	.	
$m\text{-}i_u+1$	n_{m-i_u+1}	$= a((m\text{-}i_u+1)+b)^c$

By substituting the equated formula of n_m, $n_{m-1},\ldots,$ and n_{m-i_u+1} into Equation 3.16, we derive Equation 3.17. [ ]

Equation 3.17 implies that when Zipf's law holds, the shape of the 80/20 curve in region I is influenced by five parameters, μ, N, a, b, and c.

Region II: the Middle Class

Equation 3.14 indicates that the shape of the curve in region II depends on m, μ, and N which are all constant. Thus, s_i, and d_i are decreasing for $m\text{-}i_u+2 \le i \le m\text{-}i_t$.

Region III: the Trivial Many

Equation 3.15 implies that

$$\begin{aligned}
(s_{m-i_t+1},d_{m-i_t+1}) &= \left(\frac{i_1}{\mu},\ \frac{1}{N}\sqrt{(\mu^2+i_1^2)f^2(i_1)}\ \right),\\
(s_{m-i_t+2},d_{m-i_t+2}) &= \left(\frac{i_1-1}{\mu},\ \frac{1}{N}\sqrt{(\mu^2+(i_1-1)^2)f^2(i_1-1)}\ \right),\\
(s_m,d_m) &= \left(\frac{1}{\mu},\ \frac{1}{N}\sqrt{(\mu^2+1)f^2(1)}\ \right).
\end{aligned} \tag{3.18}$$

Thus, Equation 3.15 indicates that the values m, μ, N, and f(j), $1 \leq j \leq i_\ell$ determine the shape of the curve. s_i is a decreasing function with respect to i, for $m\text{-}i_\ell+1 \leq i \leq m$. Since f(j) is a strictly decreasing function for $1 \leq j \leq i_\ell$, we see that d_i shows an increasing pattern for $m\text{-}i_\ell+1 \leq i \leq m$.

Recalling Equation 2.1, $f(n_i)$ denote frequencies of the "trivial many" in the region where $1 \leq i \leq i_\ell$ and $n_i = i$. One of the most cited laws of the trivial many is Lotka's law. By applying the indexed version of Lotka's law (Equation 2.2), we can simplify the slope-distance pairs in region III to be:

Corollary 3: If the index formulation of Lotka's law holds, then the slope-distance pairs of the curve in region III are

$$(s_i,d_i) = \left(\frac{m-i+1}{\mu}, \frac{1}{N}\sqrt{d\left(\mu^2+(m-i+1)^2\right)\left((m-i+1)^e-(m-i+2)^e\right)}\right), \tag{3.19}$$

where $i = m\text{-}i_\ell+1,\ldots,m$.

Proof: If the indexed formulation of Lotka's law holds, then $f(n_i) = d(n_i^e - n_{i+1}^e)$, for $i = 1,2,\ldots,i_\ell$. Since Equation 2.1 indicates $n_i = i$, $1 \leq i \leq i_\ell$, for $i = 1,2,\ldots,i_\ell$ we have $f(i) = d(i^e - (i+1)^e)$. By substituting the last equation into Equation 3.15 we derive equation 3.19 immediately. [ ]

Equation 3.19 implies that when Lotka's law holds, the shape of the 80/20 curve in region III is influenced by four parameters μ, N, d, and e. Furthermore, from Equation 3.18, $s_m = 1/\mu$. We will used s_m to estimate α in Sections 3.2 and 4.5.1.

3.1.3 On Burrell's Finding

With the insights gained from using index approach, we now revisit findings of Burrell (1985) and Egghe (1986) discussed in Section 3.1. To study Burrell's finding

of the inverse relationship between minimum holdings and the average circulation rate (Equation 3.1), we define two curves C′ and C″ to be those formed by $\{(x_i',\theta_i'): i = 1,2,...m'\}$ and $\{(x_i'',\theta_i''): i = 1,2,...m''\}$, respectively. As an example, in Figure 3.1 C′ refer to Kendall's data (Table 3.1a) curve and C″ refer to Bradford's data (Table 3.1b) curve. Burrell (1985) found that $x' < x''$ if $\mu' > \mu''$ and $\theta' = \theta'' = 0.8$. For example, in Table 3.1 and Figure 3.1, when $\theta' = \theta'' = 0.8$, $\mu'=4.765 > \mu''=2.409$ and $x'=0.25 < x''=0.52$.

We may study this effect by setting $\theta' = \theta'' = 0.8$ in Equation 3.10 and compute x′ - x″; however, the equation for x′ - x″ involves eight types of parameters. The following theorem about the slopes for the two curves C′ and C″ provides a simpler way to explain Burrell's finding.

Theorem 4: Suppose we denote the upper portion of the curve C′ and C″ to be C'_{upper} and C''_{upper}, respectively. If (x′,0.8) and (x″,0.8) fall in C'_{upper} and C''_{upper}, respectively, and $\mu' > \mu''$ then $x' < x''$ where

$$\begin{aligned} C'_{upper} &= \{(x'_{m'-i_0},\theta'_{m'-i_0}),\ldots,(x'_{m'},\theta'_{m'})\}, \\ C''_{upper} &= \{(x''_{m''-i_0},\theta''_{m''-i_0}),\ldots,(x''_{m''},\theta''_{m''})\}, \end{aligned} \tag{3.20}$$

and $i_0 = \min\{m' - i'_{l'} + 1,\ m'' - i''_{l''} + 1\}$.

Proof: Note that Equation 3.9 shows that the slopes for the two curves C′ and C″ are:

$$\begin{aligned} s'_1 &= \frac{n'_{m'}}{\mu'},\ s'_2 = \frac{n'_{m'-1}}{\mu'},\ldots,\ s'_{m'-1} = \frac{n'_2}{\mu'},\ s'_{m'} = \frac{n'_1}{\mu'}, \\ s''_1 &= \frac{n''_{m''}}{\mu''},\ s''_2 = \frac{n''_{m''-1}}{\mu''},\ldots,\ s''_{m''-1} = \frac{n''_2}{\mu''},\ s''_{m''} = \frac{n''_1}{\mu''}, \end{aligned} \tag{3.21}$$

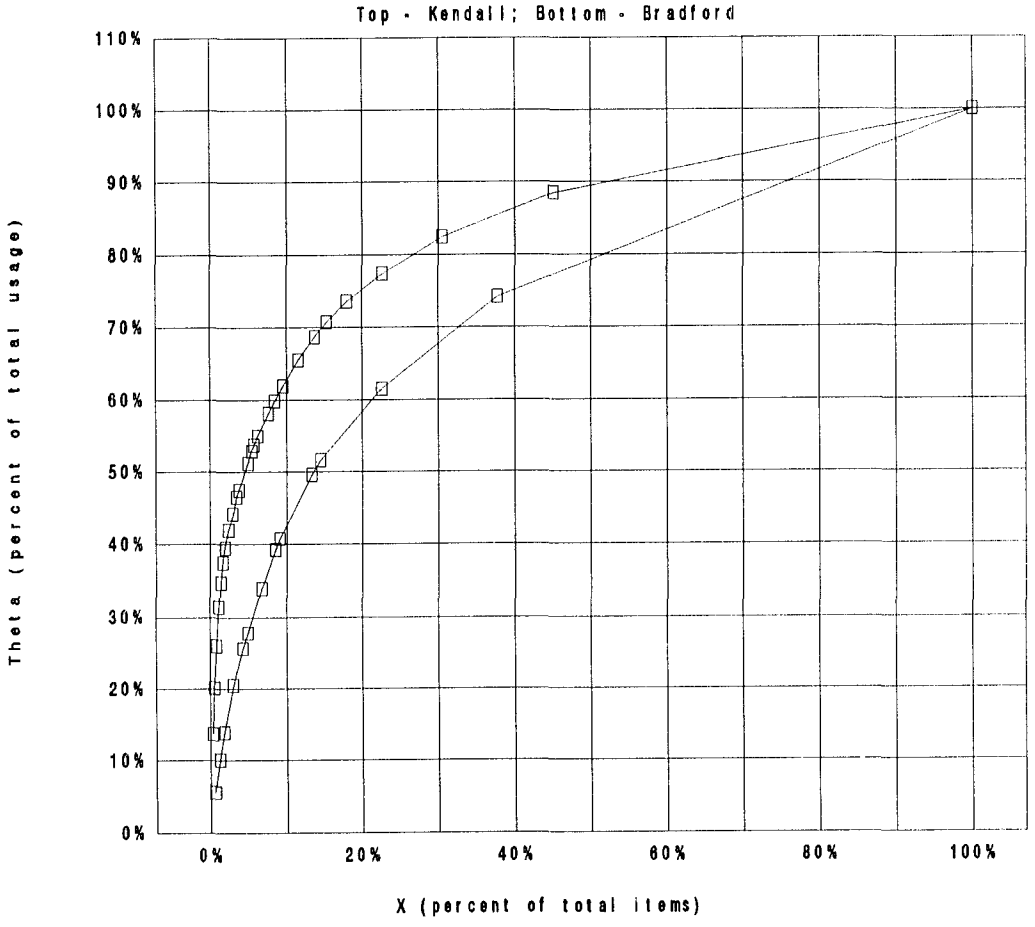


Figure 3.1: 80/20 Rule Using Kendall and Bradford's Data

Table 3.1: Indexed Data on Two Collections for 80/20 Formulations

(a) Kendall's Data (1960)

A	B	C	D	E	F
i	n_i	$f(n_i)$	$n_i f(n_i)$	x_i	θ_i
1	1	203	203	0.003	0.137
2	2	54	108	0.005	0.202
3	3	29	87	0.008	0.260
4	4	17	68	0.011	0.314
5	5	10	50	0.014	0.347
6	6	6	36	0.016	0.374
7	7	8	56	0.019	0.394
8	8	8	64	0.024	0.419
9	9	4	36	0.030	0.442
10	10	3	30	0.035	0.465
11	11	5	55	0.038	0.475
12	12	2	24	0.049	0.512
13	14	1	14	0.054	0.529
14	15	2	30	0.057	0.537
15	16	4	64	0.062	0.550
16	18	1	18	0.076	0.581
17	20	2	40	0.084	0.598
18	21	2	42	0.095	0.619
19	22	2	44	0.116	0.655
20	34	1	34	0.138	0.687
21	49	1	49	0.154	0.707
22	58	1	58	0.181	0.736
23	95	1	95	0.227	0.774
24	102	1	102	0.305	0.824
25	114	1	114	0.451	0.885
26	242	1	242	1.000	1.000
		----	-----		
m=26		T=370	N=1763	μ=4.765	

Column A = index i, i=1,2,...m.
Column B = number of papers n_i.
Column C = number of journals $f(n_i)$.
Column D = Column B * Column C.
Column E = Equation 3.4 in the text.
Column F = Equation 3.5 in the text.

Table 3.1 Continued

(b) Bradford's Data (1934)

A	B	C	D	E	F
i	n_i	$f(n_i)$	$n_i f(n_i)$	x_i	θ_i
1	1	102	102	0.006	0.056
2	2	25	50	0.012	0.101
3	3	13	39	0.018	0.139
4	4	2	8	0.030	0.205
5	5	7	35	0.043	0.256
6	6	1	6	0.049	0.278
7	7	3	21	0.067	0.339
8	8	3	24	0.085	0.392
9	9	1	9	0.091	0.408
10	10	2	20	0.134	0.496
11	13	2	26	0.146	0.516
12	15	1	15	0.226	0.615
13	18	1	18	0.378	0.742
14	22	1	22	1.000	1.000
		----	----		
m=14		T=164	N=395	μ=2.409	

Column A = index i, i=1,2,...,m.
Column B = number of papers n_i.
Column C = number of journals $f(n_i)$.
Column D = Column B * Column C.
Column E = Equation 3.4 in the text.
Column F = Equation 3.5 in the text.

which, with reference to Equation 2.1 (the three regions), can be simplified to be:

$$s'_{m'} = \frac{1}{\mu'},\ s'_{m'-1} = \frac{2}{\mu'},\ldots,\ s'_{m'-i'_{i'}+1} = \frac{i'_{i'}}{\mu'}$$
$$s''_{m''} = \frac{1}{\mu''},\ s''_{m''-1} = \frac{2}{\mu''},\ldots,\ s''_{m''-i''_{i''}+1} = \frac{i''_{i''}}{\mu''}. \qquad (3.22)$$

Equation 3.22 shows that if $\mu' > \mu''$ then $s'_{m'} < s''_{m''},\ldots, s'_{m'-i_0} < s''_{m''-i_0}$, where $i_0 = \min\{m'-i'_{i'}+1,\ m''-i''_{i''}+1\}$. Since the two curves C′ and C″ end at the same point (1,1), Equation 3.22 also implies that if $\mu' > \mu''$, then the upper portion of the curve C′ is over the upper portion of the curve C″. Thus, if (x′,0.8) and (x″,0.8) fall in C'_{upper} and C''_{upper}, respectively, and if $\mu' > \mu''$ then $x' < x''$. [ ]

As an example, Kendall and Bradford's data (Tables 3.1a and 3.1b; Figure 3.1), give m′ = 26, $i'_{i'} = 12$, m″ = 14, $i''_{i''} = 10$, and $i_0 = \min\{15,5\} = 5$. The two points (x′,0.8) and (x″,0.8) fall in $C'_{upper} = \{(x'_{21},\theta'_{21}),\ldots,(x'_{26},\theta'_{26})\}$ and $C''_{upper} = \{(x''_{9},\theta''_{9}),\ldots,(x''_{14},\theta''_{14})\}$, respectively. Hence, we may find from Figure 3.1 that $x' = 0.25 < x'' = 0.52$. Note that if one of the two points (x′,0.8) and (x″,0.8) does not lie in the upper portion of the curve C'_{upper} or C''_{upper}, then Burrell's finding might not be true. This explains the exception of Public Library B noted in Burrell's paper (1985).

3.1.4 On Egghe's Finding

In Egghe's study (1986) of the relationship between x and θ (Equation 3.2), Lotka's law was assumed. Here we use the indexed version of Lotka's law (Equation 2.2) to simplify Equation 3.10 where the important factors affecting the 80/20 rule are index size m, distribution $f(n_i)$, and the scattering pattern of n_i, for i = 1,2,…,m. Using Equation 2.2, we can derive the following theorem:

Theorem 5: Assuming the indexed version of Lotka's law holds, then

$$x \simeq n^e_{m-j+2} + \frac{1}{(d-b)n_{m-j+1}} \sum_{k=1}^{m-j+1} n_k f(n_k) - (1-\theta)\frac{\mu}{n_{m-j+1}}, \tag{3.23}$$

where θ is given and j is the smallest i such that $\theta_i \geq \theta$.

Proof: Since $F(n_i) = dn^e_i - b$ and $n_1 = 1$, we have $T = F(n_1) = d - b$,

$$\mu_j = \frac{\sum_{k=m-j+1}^{m} n_k f(n_k)}{\sum_{k=m-j+1}^{m} f(n_k)} = \frac{N - \sum_{k=1}^{m-j} n_k f(n_k)}{F(n_{m-j+1})} = \frac{\mu - \frac{\sum_{k=1}^{m-j} n_k f(n_k)}{d-b}}{x_j},$$

and, by definition,

$$x_j = \frac{F(n_{m-j+1})}{F(n_1)} = \frac{dn^e_{m-j+1} - b}{d-b} = \frac{n^e_{m-j+1} - b/d}{1 - b/d} \simeq n^e_{m-j+1}.$$

Therefore, from Equation 3.10,

$$x = \left(1 - \frac{\mu - \frac{\sum_{k=1}^{m-j+1} n_k f(n_k)}{d-b}}{x_{j-1} n_{m-j+1}}\right) x_{j-1} + \frac{\mu}{n_{m-j+1}}\theta.$$

Thus,

$$x \simeq n^e_{m-j+2} + \frac{\sum_{k=1}^{m-j+1} n_k f(n_k)}{(d-b)n_{m-j+1}} - (1-\theta)\frac{\mu}{n_{m-j+1}}. \qquad [\,]$$

Let C be the curve formed by the m points $\{(x_i,\theta_i),\ i = 1,2,\ldots m\}$. Then Equation 3.23 can be simplified to describe the upper portion of the curve by means of the following corollary.

Corollary 4: For $j = m - t$, $0 \leq t \leq i_u - 3$, where $t = m - j$ and j is the smallest i such that $x_i \geq x$:

$$x \cong \frac{1+2^e+\ldots+(t+2)^e-(1-\theta)\mu}{t+1} \tag{3.24}$$

Proof: For $j = m - t$, $0 \le t \le i_u-3$, and using Equations 2.1, 3.4, and 3.24,

$$\begin{aligned} x &\simeq n^e_{t+2} + \frac{\sum_{k=1}^{t+1} n_k f(n_k)}{(d-b)\,n_{t+1}} - (1-\theta)\frac{\mu}{n_{t+1}} \\ &= (t+2)^e + \frac{\sum_{k=1}^{t+1} k\, f(k)}{(d-b)(t+1)} - (1-\theta)\frac{\mu}{t+1} \\ &= (t+2)^e + \frac{d\sum_{k=1}^{t+1} k(n^e_k - n^e_{k+1})}{(d-b)(t+1)} - (1-\theta)\frac{\mu}{t+1} \\ &\simeq (t+2)^e + \frac{\sum_{k=1}^{t+1} k\cdot k^e - \sum_{k=1}^{t+1} k(k+1)^e}{t+1} - (1-\theta)\frac{\mu}{t+1} \end{aligned}$$ [ ]

To apply Equation 3.24 to the 80/20 rule, the point (x,0.8) must fall in the upper portion of the curve C; i.e., the point (x,0.8) is on the line segments formed by

$$C_{upper} = \{(x_{i_u-3},\theta_{i_u-3}),(x_{i_u-2},\theta_{i_u-2}),\ldots,(x_m,\theta_m)\}.$$

If this is the case, then we can plug $(x,\theta) = (0.2,0.8)$ into Equation 3.24 and obtain

$$\mu = 5[1+2^e+\ldots+(t+2)^e] - (t+1). \tag{3.25}$$

We can then conclude that the 80/20 rule holds if the parameters μ, e, and t satisfy Equation 3.25. Comparing Equations 3.10 with 3.25, we find that the indexed version of Lotka's law enables us to reduce the number of parameters influencing the 80/20 rule.

3.2 Relationship Between s_m and α in the 80/20 Rule

Equation 3.18 indicates that s_m, the slope between points (x_{m-1},θ_{m-1}) and (100%,100%), to be the inverse of the average rate of transaction, or $1/\mu$. Consider that $1/\mu$ = T/N, or the average number of distinct items per usage; since every "distinct"

item is once a "new entry" over the span of total transaction, we reason that s_m is basically the same as the probability of new entry, α.

There is a plausible reason that s_m can be used to estimate α. The last data point which leads to (100%,100%) is f(1), or the number of items which are used only once. Unlike the other items whose usages depend upon their previous usages, the make up of f(1) is strictly from Assumption I, or dependent on α. Thus, s_m approximates α. We will verify this heuristic through simulation in Section 4.5.1.

3.3 The Need of Computational Experiment

The index approach provides much insight to these empirical phenomena, and Chen (1989) has shown that Simon's generating mechanism can be used to model these empirical phenomena; however, there are limitations in what analytical studies can do. As demonstrated in this chapter, the number of parameters involved render the analytical methods impossible. On the other hand, stochastic models admitting serial correlation have proved to be too complex to be solved explicitly in closed form for the equilibrium distribution (Ijiri and Simon 1977, p. 159). As Neuts (1986b) suggested, experimentation should be used to study the validity of an hypothesis when it cannot be settled by other means. Furthermore, because the conventional analytical methods can only derive the "average behavior" of the distributions, Leimkuhler (1988) suggested that computational experimentation be used for modeling empirical phenomena, especially in studying the distributions under "extreme conditions" (Leimkuhler 1988, Neuts 1986a). Computational experimentation also allows researchers to examine details as the assumptions are relaxed (Neuts 1986b, Simon and Van Wormer 1963). Simon's three

models and their corresponding generating mechanisms will be used in Chapters 4 and 5 to show the pattern changes in these empirical phenomena as we vary the parameters.

3.4 Summary of Findings

In this chapter we take explicit account of the sequence of observed values of the variables by means of an index. As the index approach reveals, the 80/20 curve basically is influenced by the distribution of $f(n_i)$, $i = 1,2,\ldots,m$. Without making any assumption on the distribution, we are able to identify several inherent properties of the 80/20 curve as shown in Theorem 1 and Corollary 1. Continuing the analysis of the Pareto principle using the 80/20 rule, we investigate the unknown distributions $f(i)$, $1 \leq i \leq i_\ell$, and n_i, $i_u \leq i \leq m$ by posing some conditions.

Lotka's law was introduced in Corollary 3 to describe $f(i)$, $1 \leq i \leq i_\ell$; and Zipf's law was introduced in Corollary 2 to describe n_i, $i_u \leq i \leq m$. The two laws and the index approach enable us to identify the parameters influencing the three regions of the 80/20 curve. In addition:

(1) Equation 3.9, showing the slope of the 80/20-type curve, allows us to derive a sufficient condition for Burrell's inverse relationship between minimum holdings and the average circulation rate; and

(2) Equation 3.10 and the indexed version of Lotka's law enable us to derive a sufficient condition related to Egghe's finding on the 80/20 rule.

(3) Equation 3.18 provides a heuristic to estimate the probability of new entry, α, in Simon's model.

CHAPTER 4

COMPUTATIONAL RESULTS OF SIMON'S TWO BASIC MODELS

In this chapter we discuss the simulation results and show the behaviors of the 80/20 curve and other three empirical phenomena when parameters are changed in Simon's two basic models. Specifically, we try to answer the following three questions: (1) what are the effects of changing parameters? (2) what is the effect of time (expressed as the total number of usages)? and (3) under what condition would the 80/20 rule and other empirical phenomena hold? We find the rate of new entry α to have the greatest influence to the curves of these phenomena. Basically, small α generates high usage concentration. The total number of iterations (the total usages) N plays a role only when α is a decreasing probability function. Important values of α are noted, especially those under which the empirical laws hold. We also verify the heuristics suggested in Section 3.2, and we use Kendall's (1960) data to demonstrate its application.

4.1 The 80/20 Rule

Constant α

Although we can use measures such as 80/20 or 90/10 to describe the shapes of the curves which reflect the usage patterns, this unique point where the two fractions add up to 100% is not always accurately obtainable from the data. For example, even though we may approximate the measure to be 78/22 in Kendall's Data, these numbers

are found through visual inspection of the graph and do not always have a corresponding data point.

Since usage patterns with higher concentration tend to have larger area above the midline, we devise a parameter Area to measure the level of usage concentration. The parameter Area is defined as the area between the curve formed by $\{(x_i,\theta_i),\ i=1,2,\ldots,m\}$ and the $\theta_i = x_i$ line, i.e.,

$$\begin{aligned} \text{Area} &= \frac{\theta_1 x_1}{2} + \frac{(\theta_1+\theta_2)(x_2-x_1)}{2} + \frac{(\theta_2+\theta_3)(x_3-x_2)}{2} + \ldots + \frac{(\theta_{m-1}+\theta_m)(x_m-x_{m-1})}{2} - \frac{1}{2} \\ &= \frac{1}{2}\left[(\theta_1 x_2 - \theta_2 x_1) + (\theta_2 x_3 - \theta_3 x_2) + \ldots + (\theta_{m-1} x_m - \theta_m x_{m-1}) - 1\right] \end{aligned} \tag{4.1}$$

Thus, an Area of 0 means that $x_i = \theta_i$ everywhere; each item is used only once, and there is no concentration whatsoever.

Table 4.1 is a summary of the simulation results of seven different parameters α, N, m, n_m, f(1), μ, and Area. It shows that the N does not seem to affect the outcome of the simulation results. The simulation was carried out with N ranging from 1,000 to 30,000 usages, and the resulting values of Area and μ were strikingly similar at all usage levels. For example, when $\alpha = 0.20$, Area varies within a narrow range of 0.3561 and 0.3593, and μ fluctuates between 4.9538 to 5.4645. We can conclude that when other things held equal, the total number of usage N does not change the usage pattern.

On the other hand, when N is held constant, say at 20,000, results show that Area and α are inversely related. Visually, smaller α produces graphs that curve to the northwest direction (Figure 4.1), and numerically, Figure 4.2 illustrates this relationship, using N = 20,000. Since the graph is near linear, a simple regression analysis is employed to obtain the relationship Area = $0.4612 - 0.4725\alpha$ with $R^2 = 0.9917$.

Table 4.1: 80/20 Rule Simulation Results: Constant Entry Rates

α	N (000)	m	n_m	f(1)	μ	Area
0.01	1	6	534	4	111.1100	0.3840
	5	13	2624	25	113.6300	0.4711
	10	16	5235	54	95.2380	0.4827
	15	19	7917	70	100.6711	0.4854
	20	23	10578	98	97.5610	0.4871
	25	27	13216	123	101.6200	0.4882
	30	28	15740	153	98.6842	0.4889
0.10	1	16	371	47	11.3636	0.4184
	5	34	1618	254	10.5485	0.4262
	10	46	3015	540	10.0200	0.4245
	15	57	4383	791	9.9734	0.4240
	20	66	5722	1025	10.1420	0.4245
	25	81	2332	2738	10.2669	0.3591
	30	84	2677	3272	10.2916	0.3589
0.18	1	23	181	80	5.9880	0.3655
	5	41	697	486	5.6180	0.3715
	10	54	1257	986	5.5157	0.3702
	15	64	1765	1525	5.4526	0.3703
	20	72	2246	1982	5.5203	0.3711
	25	78	2693	2484	5.5371	0.3712
	30	86	3093	2958	5.5607	0.3714
0.20	1	22	175	90	5.4645	0.3561
	5	43	642	537	5.0505	0.3593
	10	55	1120	1105	4.9554	0.3582
	15	65	1547	1682	4.9358	0.3582
	20	74	1935	2170	5.0112	0.3589
	25	81	2332	2738	5.0070	0.3591
	30	84	2677	3272	5.0100	0.3589
0.50	1	17	35	319	2.0161	0.2077
	5	27	67	1651	2.0105	0.2107
	10	38	94	3242	2.0346	0.2132
	15	43	119	4837	2.0305	0.2122
	20	48	137	6576	2.0169	0.2118
	25	50	148	8235	2.0134	0.2114
	30	55	161	9857	2.0146	0.2114
0.70	1	10	13	556	1.3966	0.1211
	5	15	19	2625	1.4492	0.1314
	10	19	25	5305	1.4422	0.1304
	15	20	28	8096	1.4310	0.1288
	20	23	31	10788	1.4323	0.1292
	25	23	32	13500	1.4310	0.1289
	30	25	32	16218	1.4286	0.1283
0.99	1	3	3	975	1.0131	0.0064
	5	3	3	4916	1.0086	0.0042
	10	3	3	9838	1.0082	0.0040
	15	3	3	14744	1.0086	0.0042
	20	3	3	19635	1.0092	0.0045
	25	3	3	24544	1.0092	0.0045
	30	3	3	29450	1.0093	0.0045

R = total number of usages,
α = entry rate of new items,
m = the maximal index as defined in Section 2.1,
n_m = the maximal number of usage for an item,
f(1) = the number of items which have been used 1 time,
μ = average usage per item, and
Area = Equation 4.1 in the text.

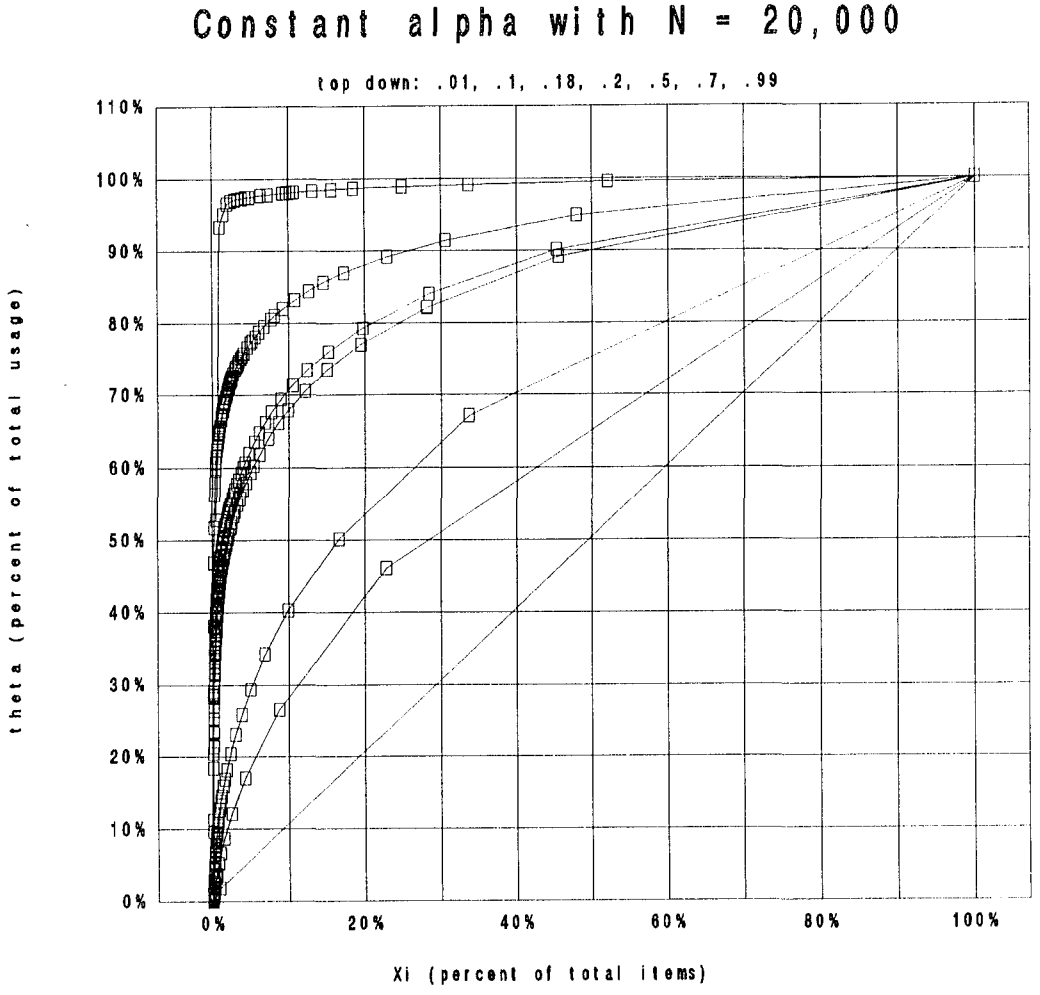


Figure 4.1: Results for 80/20 Formulation, Constant α

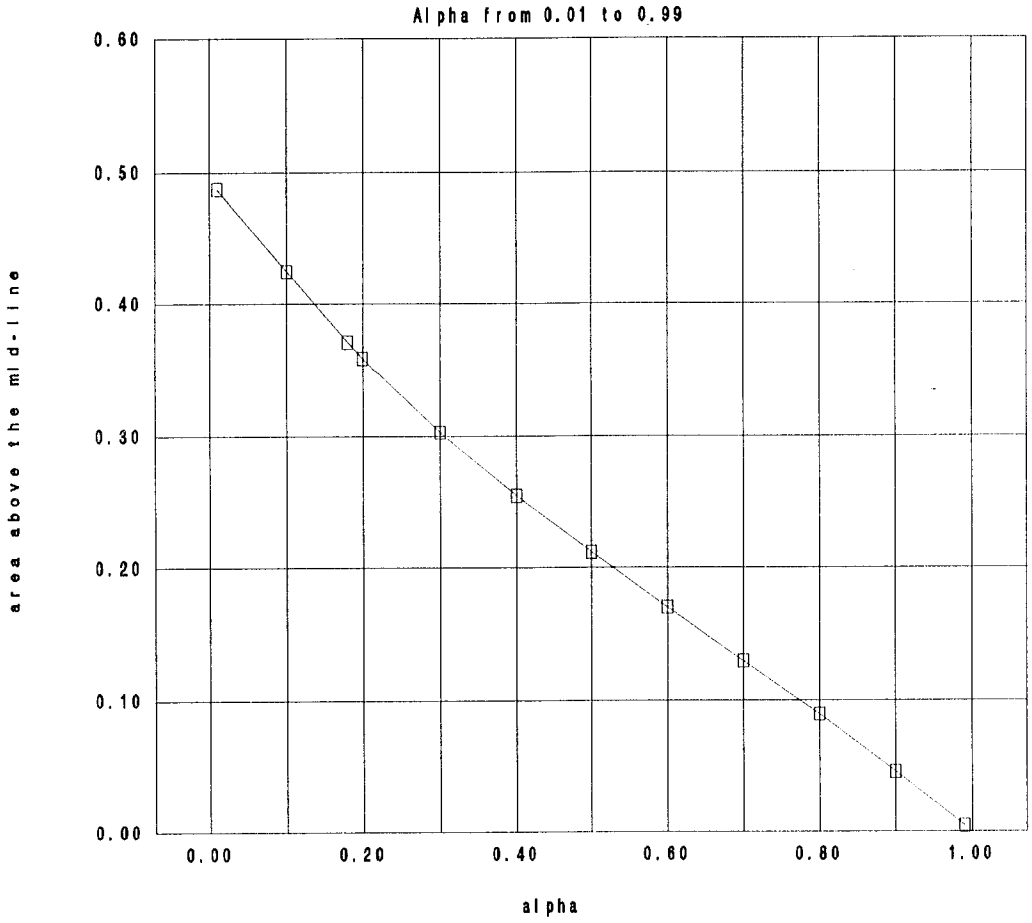


Figure 4.2: Areas Under 80/20 Curve (N = 20,000)

The inverse relationship implies that an increase in the probability of new entries (higher α) increases the number of distinct items accessed. The usages are spread over more items, resulting in less concentration in usage patterns. This is evident in the reduction of n_m, the usage frequency of the most-used item. For example, in Table 4.1, with N = 20,000, n_m declines from 10,578 (α = 0.01) to 13 (α = 0.70) — that is, nothing is used more than 13 times when α = 0.70.

Since the relationship between Area and α is smooth and linear, Figure 4.1 indicates that 80/20 is reached only when $\alpha \approx .18$; however, for the convenience of discussion we will refer to this α as 0.20.

Decreasing α

Several decreasing functions were used in the simulations, and results are summarized in Table 4.2. In general, results show that the faster the α decreases, the higher the concentration (Figure 4.3). For example, let $\alpha(R) = A/\ln(R)$, R = 1,2,...,N, where A is a constant ranging from 1 to 2 with 0.25 increment, and N = 20,000, the measure of 80/20 can be approximated at A = 1.25. When $\alpha(R) = 2/\ln(R)$ the measure is approximately 75/25.The regression analysis obtains Area = 0.4750 - 0.0748A, with R^2 = 0.9977. Thus, faster decreasing functions generate lower α which in term translates to lower probability of new entries, resulting in higher concentrations. On the other hand, higher N increase usage concentration by reducing α in the long run.

Our simulation results indicate that α affects the usage concentration in both cases of constant and decreasing α. We will discuss the method of estimating α of an empirical data through analyzing its 80/20 curve in Section 4.5.

Table 4.2: 80/20 Rule Simulation Results at Decreasing Entry Rates $\alpha = A/\ln(R)$, $R = 1,2,\ldots N$.

A	N(000)	m	n_m	f(1)	μ	Area
1.00	1	26	101	70	5.9172	0.3422
	5	48	426	346	7.2886	0.3821
	10	68	794	630	7.9239	0.3921
	15	80	1143	881	8.3333	0.3971
	20	84	1500	1098	8.7719	0.4014
	25	95	1854	1323	9.1008	0.4042
	30	101	2192	1579	9.3168	0.4066
1.25	1	24	62	95	4.7169	0.3135
	5	55	254	417	5.9171	0.3582
	10	68	476	789	6.3452	0.3698
	15	80	675	1097	6.7355	0.3763
	20	96	879	1413	6.9905	0.3811
	25	100	1074	1706	7.2380	0.3846
	30	111	1260	2014	7.4129	0.3872
1.50	1	23	51	139	3.6900	0.2925
	5	48	194	514	4.8402	0.3349
	10	68	365	939	5.2938	0.3482
	15	80	515	1365	5.5555	0.3563
	20	95	673	1733	5.8055	0.3616
	25	102	800	2098	5.9895	0.3653
	30	111	955	2470	6.1287	0.3682
1.75	1	22	39	165	3.1545	0.2671
	5	49	133	620	4.1017	0.3144
	10	67	252	1137	4.4743	0.3289
	15	80	356	1657	4.6904	0.3371
	20	88	460	2072	4.9152	0.3427
	25	104	539	2512	5.0813	0.3468
	30	110	641	2937	5.2047	0.3497
2.00	1	21	28	185	2.8735	0.2501
	5	42	97	720	3.6576	0.2994
	10	62	182	1331	3.9494	0.3140
	15	77	248	1875	4.1666	0.3219
	20	89	314	2347	4.3610	0.3271
	25	100	362	2893	4.4779	0.3315
	30	109	432	3365	4.5837	0.3343

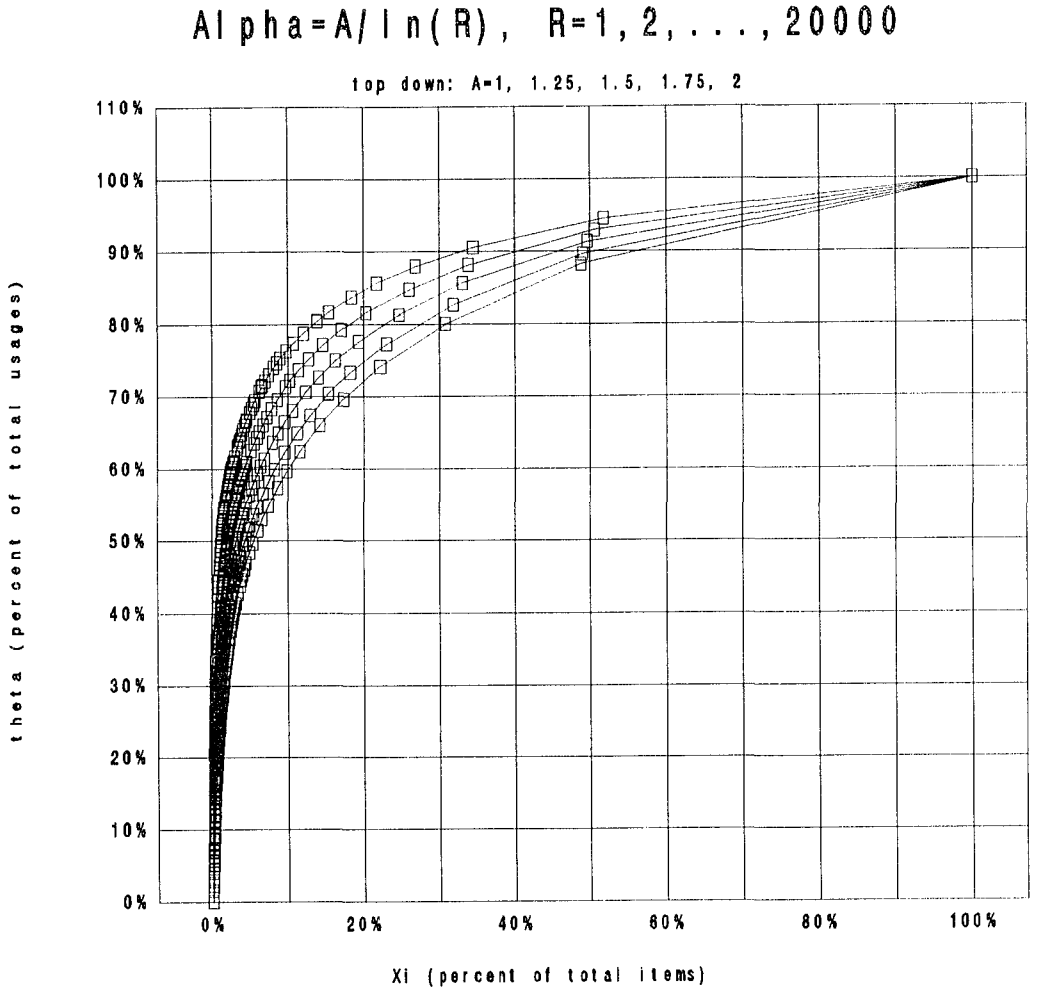


Figure 4.3: Results for 80/20 Formulation, Decreasing α

4.2 Lotka's Law

Since a full graph of Lotka's law in most of our simulation results only show curves hugging both axes, we only show the partial graph near the origin in our figures. In addition to the original presentation of Lotka's law of using n_i and $f(n_i)$; we use $\log(n_i)$ and $\log(f(n_i))$ to highlight the effects of changing parameters.

Constant α

Figure 4.4 is one example of the simulation result of Lotka's curves, with N = 20,000 and α = 0.10 and 0.90. In general, as indicated in the previous section, a large α tends to have a smaller n_m (the horizontal part of the graph), and it reduces the curve to a near vertical line approaching the y-axis. In terms of the three clusters described in Section 2.1.2, high α decreases the cluster 3 where $f(n_i) = 1$. Note that when $\alpha \approx 0.90$, $n_m = m$ and there is no excessive cluster 3. In other words, with high α, we can find items that have been used n times, with n running from 1 to m, consecutively.

In order to describe the changing curvatures of these graphs, we define the parameter Area_L to be the area under the Lotka's curve, formed by $\{(n_i, f(n_i)), i=1,2,\ldots,m\}$, i.e.,

$$\text{Area}_L = \frac{1}{2}[(f(n_1)+f(n_2))(n_2-n_1) + (f(n_2)+f(n_3))(n_3-n_2) + \ldots + (f(n_{m-1}) + f(n_m))(n_m-n_{m-1})] \quad (4.2)$$

Table 4.3 shows that Area_L increases linearly with respect to the size of N when α is held equal. For example, at α = 0.01, Area_L = 15943.0 when N = 30,000 — or approximately 30 times of the Area_L of 536 when N = 1,000. Since N has little effect on the shape of the curves, we arbitrarily selected N = 20,000 as a representative in our

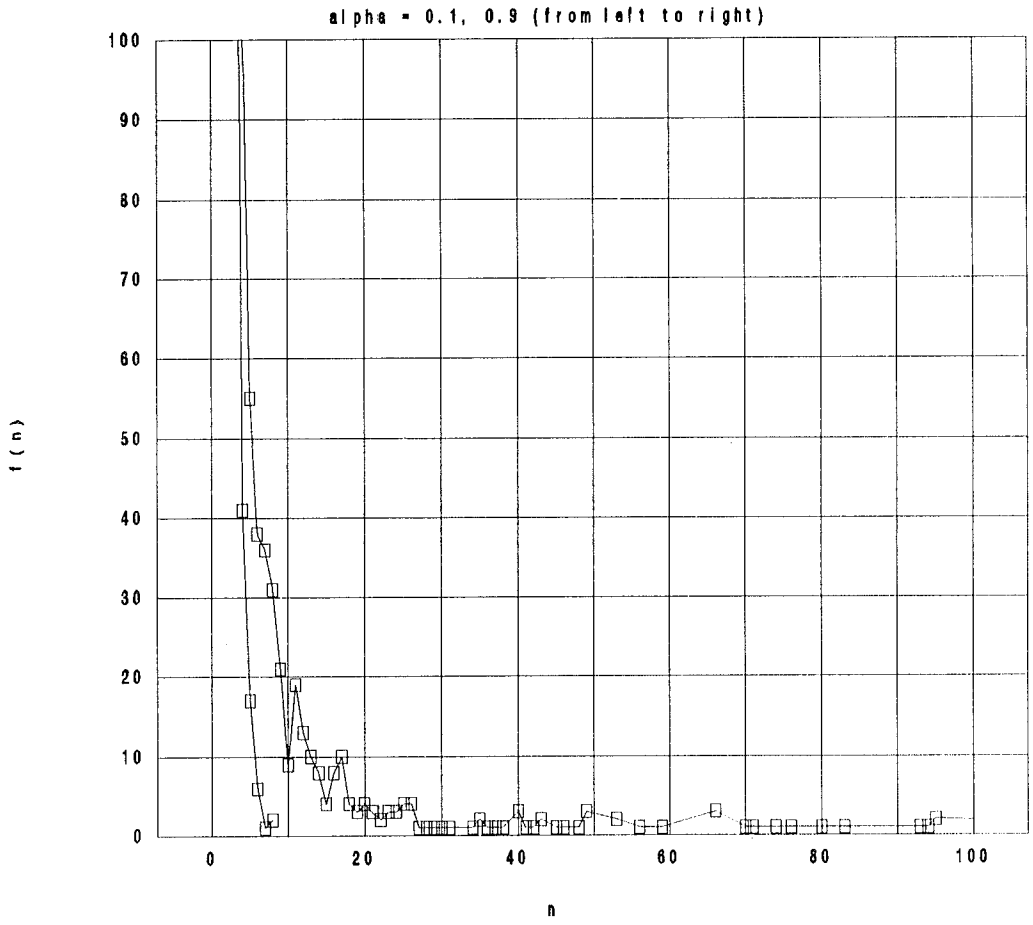


Figure 4.4: Results for Lotka's Law, Constant α

Table 4.3: Simulation Results of Lotka's Law and Zipf's Second Law, Constant α

α	N (000)	m	n_m	T	f(1)/ T	f(2)/ f(1)	f(3)/ f(1)	f(4)/ f(1)	f(5)/ f(1)	$Area_L$	A_L
0.01	1	6	534	9	0.4444	0.0000	0.2500	0.0000	0.0000	536	0.2509
0.10	1	16	371	88	0.5341	0.2128	0.1702	0.1915	0.0638	420	0.0241
0.20	1	22	175	183	0.4918	0.4556	0.1556	0.0556	0.0556	292	0.0185
0.30	1	21	91	302	0.5993	0.2099	0.1381	0.1105	0.0497	282	0.0171
0.40	1	18	44	401	0.6060	0.2510	0.1728	0.0658	0.0412	316	0.0295
0.50	1	17	35	496	0.6431	0.2978	0.0909	0.0596	0.0251	356	0.0319
0.60	1	13	16	611	0.7201	0.2114	0.0818	0.0341	0.0136	394	0.0559
0.70	1	10	13	716	0.7765	0.1871	0.0486	0.0306	0.0018	441	0.0609
0.80	1	6	8	799	0.8260	0.1530	0.0364	0.0152	0.0030	472	0.0894
0.90	1	5	5	906	0.9205	0.1367	0.0360	0.0144	0.0120	488	0.1170
0.99	1	3	3	987	0.9878	0.0113	0.0010	0.0000	0.0000	499	0.1706
0.01	5	13	2624	44	0.5682	0.2000	0.0800	0.1200	0.0400	2642	0.0403
0.10	5	34	1618	474	0.5359	0.2913	0.1654	0.0748	0.0591	1935	0.0047
0.20	5	43	642	990	0.5424	0.3240	0.1620	0.0950	0.0447	1331	0.0039
0.30	5	40	303	1523	0.5896	0.3129	0.1102	0.0735	0.0367	1339	0.0049
0.40	5	39	109	2014	0.6346	0.2551	0.1072	0.0689	0.0352	1450	0.0104
0.50	5	27	67	2487	0.6639	0.2489	0.1127	0.0424	0.0279	1705	0.0154
0.60	5	21	26	2958	0.7093	0.2288	0.0791	0.0381	0.0186	1918	0.0352
0.70	5	15	19	3450	0.7609	0.1996	0.0530	0.0232	0.0175	2142	0.0429
0.80	5	9	12	3954	0.8283	0.1450	0.0345	0.0168	0.0046	2321	0.0590
0.90	5	7	7	4492	0.9103	0.0792	0.0147	0.0037	0.0005	2447	0.0855
0.99	5	3	3	4957	0.9917	0.0079	0.0004	0.0000	0.0000	2498	0.1694
0.01	10	16	5235	105	0.5143	0.3333	0.1481	0.0741	0.0926	5300	0.0187
0.10	10	46	3015	998	0.5411	0.3019	0.1204	0.0963	0.0667	3704	0.0023
0.20	10	55	1120	2018	0.5476	0.3113	0.1620	0.0887	0.0516	2545	0.0021
0.30	10	52	506	3036	0.5827	0.3041	0.1357	0.0820	0.0447	2626	0.0029
0.40	10	49	174	3985	0.6211	0.2764	0.1329	0.0549	0.0303	2879	0.0067
0.50	10	38	94	4915	0.6596	0.2619	0.0956	0.0506	0.0312	3353	0.0110
0.60	10	27	36	5933	0.7106	0.2265	0.0804	0.0372	0.0209	3838	0.0253
0.70	10	19	25	6934	0.7651	0.1919	0.0586	0.0234	0.0124	4287	0.0323
0.80	10	10	12	7939	0.8293	0.1472	0.0328	0.0150	0.0050	4649	0.0588
0.90	10	7	7	9008	0.9112	0.0803	0.0125	0.0035	0.0007	4904	0.0853
0.99	10	3	3	9918	0.9919	0.0079	0.0002	0.0000	0.0000	4998	0.1693
0.01	15	19	7917	149	0.4698	0.3429	0.2857	0.0857	0.0714	8034	0.0145
0.10	15	57	4383	1504	0.5259	0.3338	0.1517	0.0796	0.0544	5446	0.0016
0.20	15	65	1547	3039	0.5535	0.3002	0.1576	0.0779	0.0559	3703	0.0014
0.30	15	58	649	4538	0.5837	0.2990	0.1351	0.0774	0.0521	3838	0.0022
0.40	15	56	220	5930	0.6147	0.2914	0.1265	0.0642	0.0326	4285	0.0053
0.50	15	43	119	7388	0.6547	0.2756	0.0980	0.0476	0.0269	5056	0.0088
0.60	15	30	46	8917	0.7102	0.2291	0.0764	0.0403	0.0210	5777	0.0198
0.70	15	20	28	10482	0.7724	0.1821	0.0581	0.0240	0.0120	6446	0.0284
0.80	15	12	15	11954	0.8329	0.1443	0.0318	0.0131	0.0061	6980	0.0467
0.90	15	8	8	13491	0.9106	0.0801	0.0136	0.0031	0.0010	7348	0.0748
0.99	15	3	3	14871	0.9915	0.0085	0.0001	0.0000	0.0000	7498	0.1695

Table 4.3 Continued

α	N (000)	m	n_m	T	f(1)/ T	f(2)/ f(1)	f(3)/ f(1)	f(4)/ f(1)	f(5)/ f(1)	$Area_L$	A_L
0.01	20	23	10578	205	0.4780	0.3878	0.1837	0.1327	0.0612	10713	0.0103
0.10	20	66	5722	1972	0.5198	0.3356	0.1483	0.1063	0.0537	7162	0.0012
0.20	20	74	1935	3991	0.5437	0.3217	0.1581	0.0820	0.0539	4780	0.0011
0.30	20	69	783	6023	0.5816	0.3060	0.1285	0.0748	0.0514	4997	0.0018
0.40	20	60	260	8011	0.6243	0.2729	0.1244	0.0630	0.0338	5719	0.0044
0.50	20	48	137	9916	0.6632	0.2576	0.0999	0.0465	0.0295	6719	0.0075
0.60	20	34	48	11899	0.7131	0.2209	0.0844	0.0346	0.0206	7674	0.0188
0.70	20	23	31	13963	0.7726	0.1813	0.0586	0.0227	0.0125	8578	0.0256
0.80	20	12	15	15936	0.8328	0.1435	0.0333	0.0129	0.0057	9309	0.0468
0.90	20	8	8	18015	0.9112	0.0804	0.0129	0.0025	0.0010	9806	0.0747
0.99	20	3	3	19816	0.9909	0.0091	0.0002	0.0000	0.0000	9997	0.1697
0.01	25	27	13216	246	0.5000	0.3008	0.2033	0.1057	0.0650	13376	0.0082
0.10	25	72	6961	2435	0.5170	0.3249	0.1581	0.1033	0.0627	8710	0.0010
0.20	25	81	2332	4993	0.5484	0.3112	0.1494	0.0829	0.0606	5883	0.0009
0.30	25	75	917	7539	0.5832	0.2986	0.1321	0.0771	0.0484	6202	0.0015
0.40	25	66	295	9992	0.6247	0.2695	0.1245	0.0628	0.0344	7114	0.0039
0.50	25	50	148	12417	0.6632	0.2590	0.0938	0.0488	0.0270	8404	0.0069
0.60	25	37	50	14892	0.7123	0.2252	0.0802	0.0347	0.0228	9606	0.0181
0.70	25	23	32	17470	0.7728	0.1810	0.0581	0.0233	0.0132	10731	0.0248
0.80	25	13	16	19958	0.8342	0.1412	0.0348	0.0121	0.0055	11637	0.0437
0.90	25	8	8	22509	0.9109	0.0803	0.0134	0.0027	0.0009	12256	0.0747
0.99	25	3	3	24770	0.9909	0.0090	0.0002	0.0000	0.0000	12496	0.1697
0.01	30	28	15740	304	0.5033	0.3203	0.2288	0.0784	0.0523	15943	0.0066
0.10	30	84	8154	2915	0.5252	0.2972	0.1450	0.1156	0.0614	10224	0.0008
0.20	30	84	2677	5988	0.5464	0.3139	0.1494	0.0889	0.0581	6964	0.0008
0.30	30	83	1031	9033	0.5818	0.2986	0.1376	0.0735	0.0464	7370	0.0014
0.40	30	68	327	11987	0.6243	0.2700	0.1235	0.0643	0.0339	8529	0.0035
0.50	30	55	161	14891	0.6619	0.2605	0.0989	0.0518	0.0264	10071	0.0063
0.60	30	38	51	17899	0.7122	0.2255	0.0813	0.0346	0.0223	11545	0.0178
0.70	30	25	32	21000	0.7723	0.1839	0.0572	0.0231	0.0118	12898	0.0249
0.80	30	13	16	23988	0.8348	0.1416	0.0334	0.0123	0.0055	13980	0.0436
0.90	30	8	8	27036	0.9112	0.0804	0.0131	0.0026	0.0009	14717	0.0747
0.99	30	3	3	29722	0.9908	0.0090	0.0002	0.0000	0.0000	14994	0.1697

discussion. We select a slightly higher N also to avoid the instability that usually occurs when the number of usage is low.

Since large N automatically increase $Area_L$, $Area_L$ is adjusted as a fraction of the corresponding $n_m f(1)$ — the rectangular area that has the two maximum values at its corners. The resulting fractions are denoted A_L, and they are listed in Table 4.3. Figure 4.5 indicates how A_L varies at different levels of α and N. It is clear that regardless the magnitude of N, larger α increases A_L. A_L continues to increase with α and finally converges to approximately 16.97% when $\alpha = 0.99$ for all N. The only exception takes place at the other extreme condition $\alpha = 0.01$ where A_L is higher than when $\alpha = 0.1$.

This convergence is characterized by the uniform $n_m = 3$ for all N when $\alpha = 0.99$, meaning that due to the high probability of new entry no one item is used more than 3 times. From Equation 4.2

$$Area_L = \frac{f(1)+f(2)}{2} + \frac{f(2)+f(3)}{2} = \frac{f(1)+2f(2)+f(3)}{2}.$$

Since we have defined $A_L = Area_L/n_m f(1)$, and $n_m = 3$ for all N, we have

$$A_L = \frac{f(1)+2f(2)+f(3)}{6f(1)} = \frac{1}{6} + \frac{2f(2)+f(3)}{6f(1)}.$$

With 1/6 = 16.67% and f(1) quite large (see Table 4.3), the convergence to approximately 16.9% is a logical conclusion.

The discrepancy between m and n_m in Table 4.3 indicates the nature of scattering observed values in the dataset. For example, the first row in Table 4.3 (with $\alpha = 0.01$ and N = 1,000) shows $n_m = 534$ and m = 6.

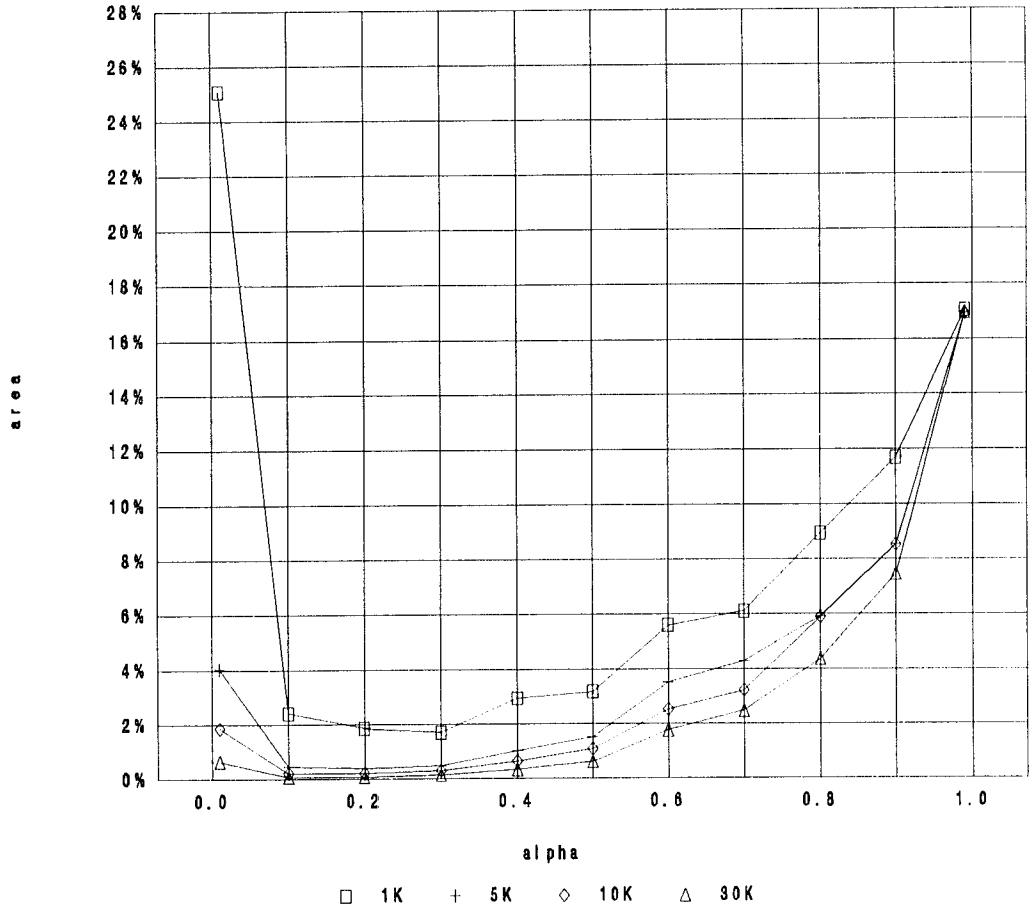


Figure 4.5: Area under Lotka's Curve (A_L), Constant α

We also use a modified Lotka's curve to show the effect of α. Figures 4.6a through 4.6d (α = 0.1, 0.3, 0.5, and 0.9, respectively) shows Lotka's curves using $\log(n_i)$ and $\log(f(n_i))$ to highlight the rate of change. These four graphs show how the slope of the curve and the three clusters are affected by α. Since cluster 3 is characterized by $f(n_i) = 1$, it is the section where $\log(f(n_i)) = 0$. On the other hand, cluster 1 (where $n_i = i$) is characterized by its linearity. Thus, Figure 4.6 shows that when α is high: (1) cluster 3 shrinks to eventually nothing; and (2) the slope becomes steeper, basically through increasing the number of low-usage items.

We find that when α = 0.30 (Figure 4.6b) the slope of this modified Lotka's curve is estimated to be approximately -2 — by which Lotka's law holds. Consistent with Lotka's prediction, f(1) is nearly 60% of the total items T at this α.

Decreasing α

Simulation results are summarized in Table 4.4. Similar to the effect of constant α, increasing A causes the curve to move away from the origin. However, Table 4.4 indicates that although $Area_L$ increases with α at a slower rate than in the case of constant α. Furthermore, when N is large, the rate of increase is even smaller. This is understandable since eventually the decreasing function will generate a small enough α (when R becomes large enough) that the value of A has little effect to the value of the function. Thus, the total number of usage N indirectly affects usage pattern by directly affecting α. Consistent with these observations, Table 4.4 shows that the decreasing A_L with respect to N is caused largely by the much faster increase in $n_m f(1)$.

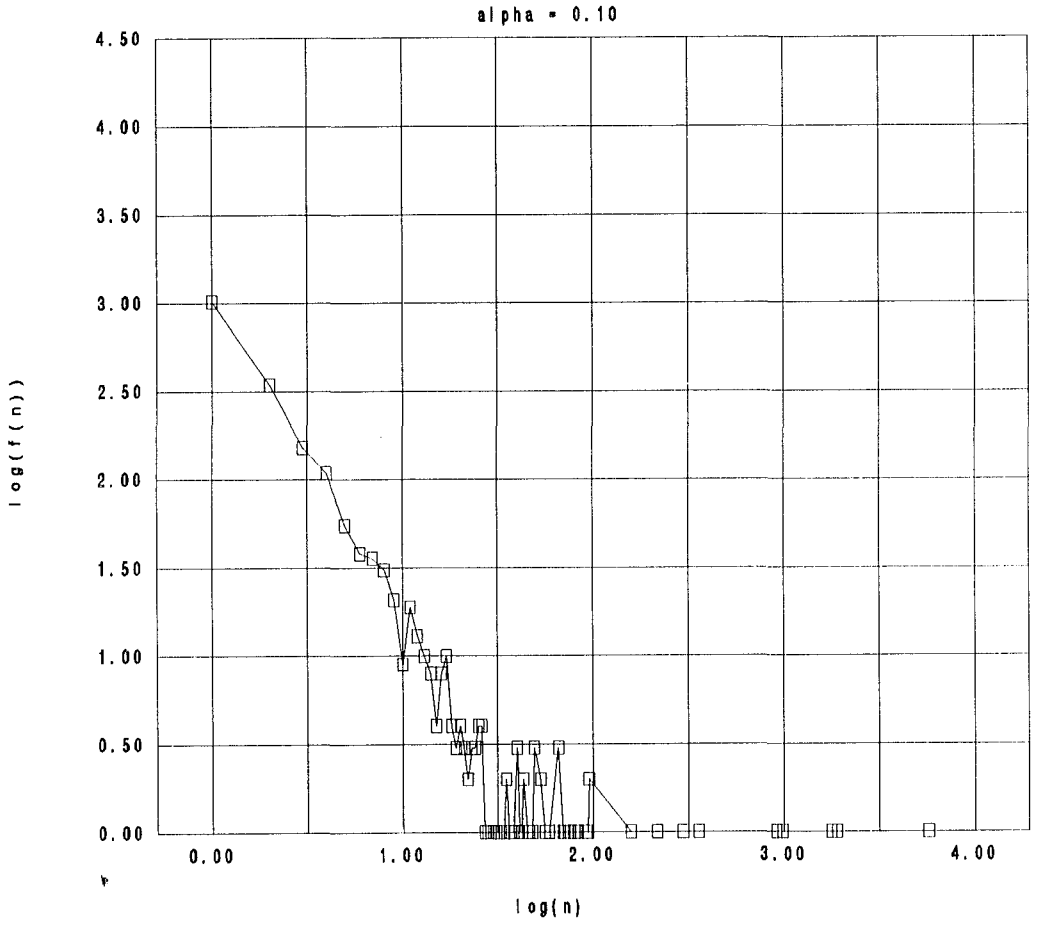


Figure 4.6a: Modified Lotka's Curve, $\alpha = 0.10$

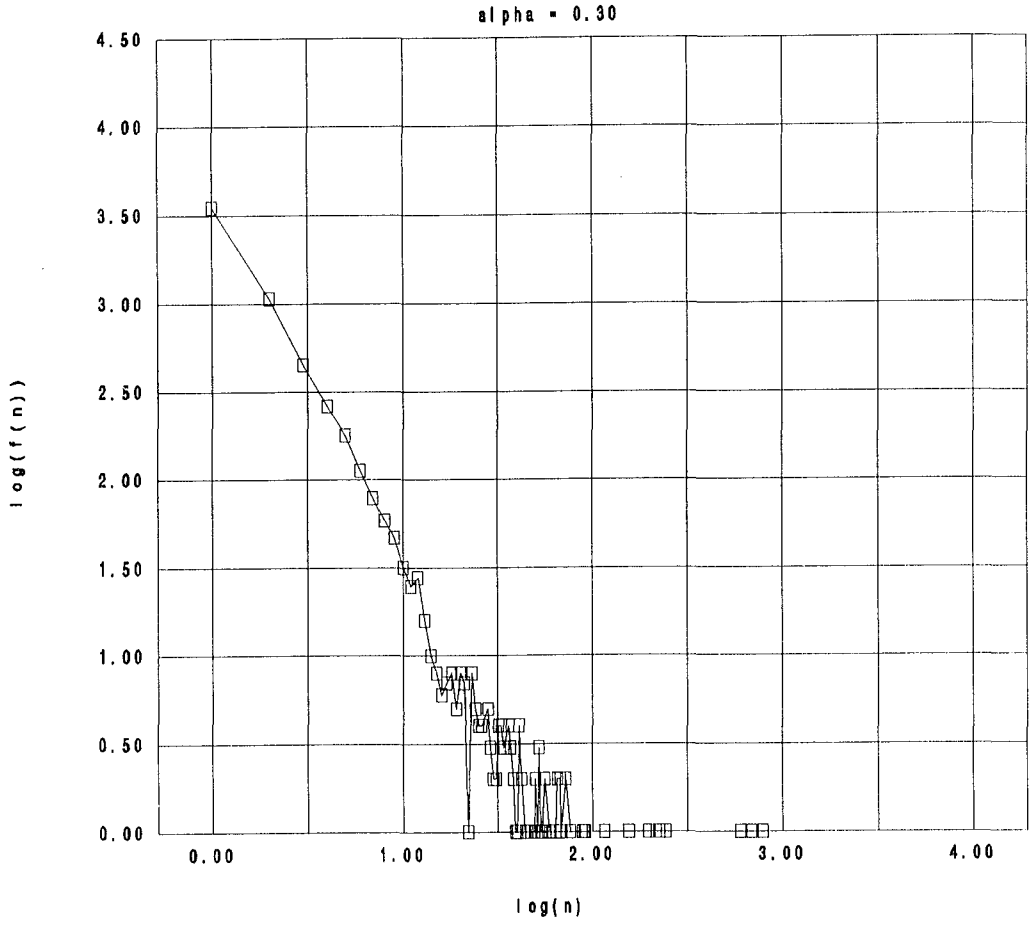


Figure 4.6b: Modified Lotka's Curve, $\alpha = 0.30$

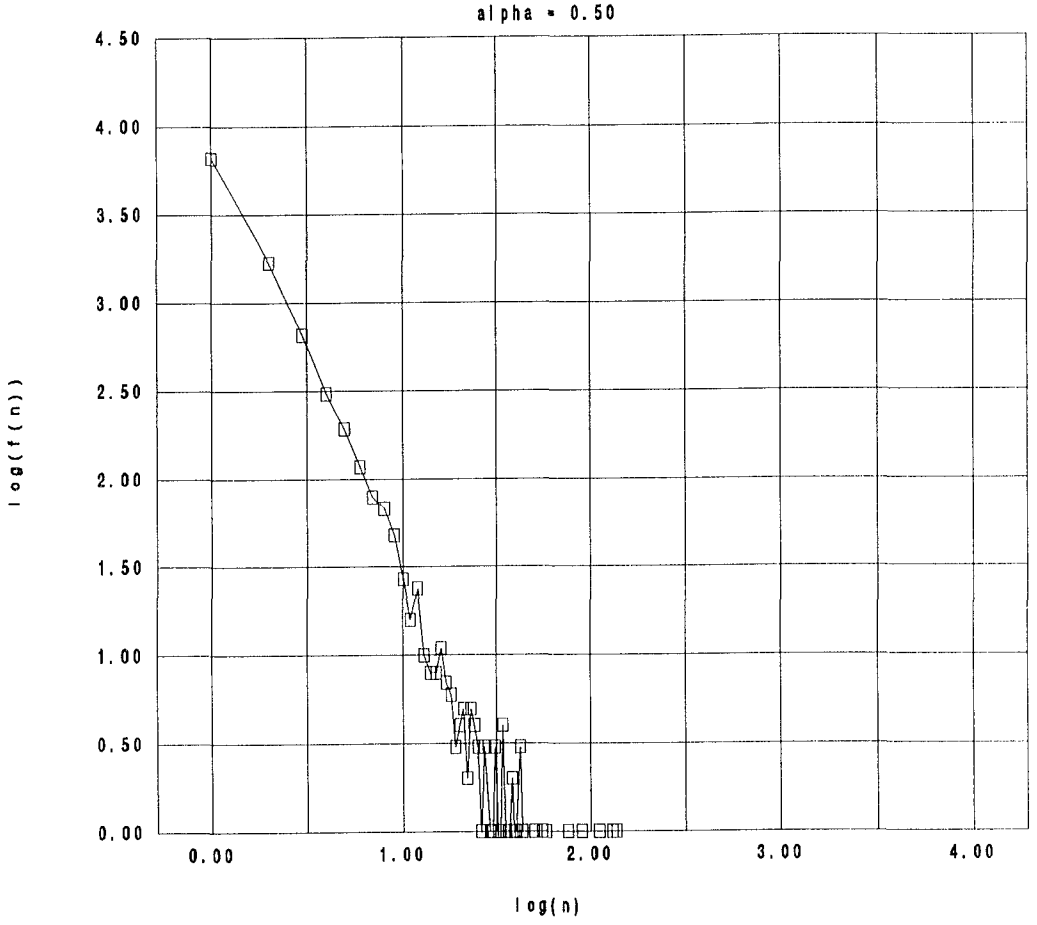


Figure 4.6c: Modified Lotka's Curve, $\alpha = 0.50$

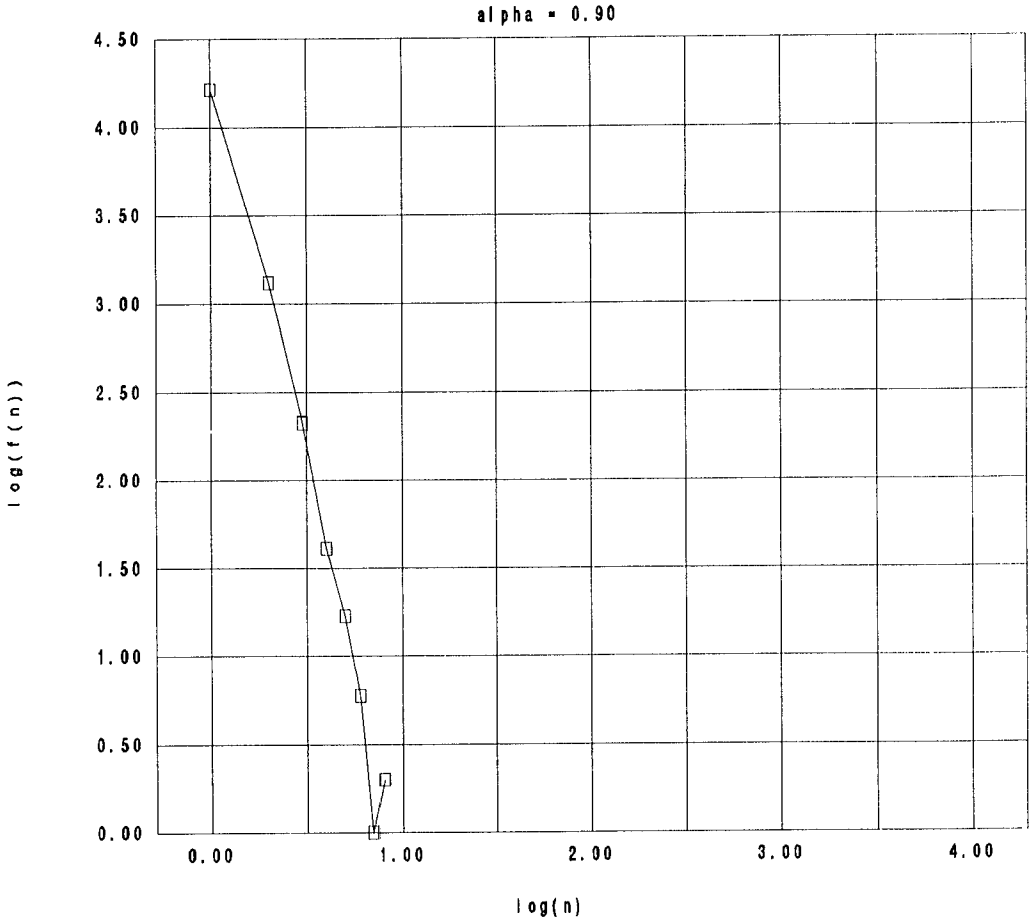


Figure 4.6d: Modified Lotka's Curve, $\alpha = 0.9$

Table 4.4: Simulation Results of Lotka's Law, Decreasing α, α = A/ln(R) where R = 1,2,...,N

A	N (000)	m	n_m	T	f(1) /T	f(2) /f(1)	f(3) /f(1)	f(4) /f(1)	f(5) /f(1)	$Area_L$	A_L
1.00	1	26	101	169	0.4142	0.6143	0.1571	0.1000	0.0429	214.5	0.0303
1.25	1	24	62	212	0.4481	0.4421	0.1895	0.0526	0.0421	204.0	0.0346
1.50	1	23	51	271	0.5129	0.3309	0.1079	0.1079	0.1079	230.5	0.0325
1.75	1	22	39	317	0.5205	0.3091	0.1394	0.1818	0.0727	257.0	0.0399
2.00	1	21	28	348	0.5316	0.2811	0.1676	0.1622	0.0595	264.5	0.0511
1.00	5	48	426	686	0.5044	0.2977	0.1590	0.0838	0.0694	917.0	0.0062
1.25	5	55	254	845	0.4935	0.3309	0.1511	0.1199	0.0959	838.0	0.0079
1.50	5	48	194	1033	0.4976	0.3346	0.1868	0.1109	0.0603	928.0	0.0093
1.75	5	49	133	1219	0.5086	0.3387	0.1935	0.0952	0.0645	994.0	0.0121
2.00	5	42	97	1367	0.5267	0.3417	0.1528	0.0931	0.0569	1076.0	0.0154
1.00	10	68	794	1262	0.4992	0.2952	0.1714	0.0984	0.0810	1676.5	0.0034
1.25	10	68	476	1576	0.5006	0.3105	0.1736	0.1065	0.0659	1606.0	0.0043
1.50	10	68	365	1889	0.4971	0.3365	0.1715	0.1076	0.0756	1724.5	0.0050
1.75	10	67	252	2235	0.5087	0.3369	0.1803	0.0915	0.0730	1853.5	0.0065
2.00	10	62	182	2532	0.5257	0.3366	0.1630	0.0924	0.0594	1996.0	0.0082
1.00	15	80	1143	1800	0.4894	0.3326	0.1657	0.0988	0.0795	2427.0	0.0024
1.25	15	80	675	2227	0.4926	0.3254	0.1778	0.1112	0.0720	2298.5	0.0031
1.50	15	80	515	2700	0.5056	0.3216	0.1692	0.1070	0.0571	2471.5	0.0035
1.75	15	80	356	3198	0.5181	0.3132	0.1768	0.0863	0.0670	2654.0	0.0045
2.00	15	77	248	3600	0.5208	0.3419	0.1552	0.1003	0.0624	2844.0	0.0061
1.00	20	84	1500	2280	0.4816	0.3597	0.1585	0.1056	0.0692	3176.5	0.0019
1.25	20	96	879	2861	0.4939	0.3390	0.1599	0.1125	0.0672	2964.0	0.0024
1.50	20	95	673	3445	0.5030	0.3289	0.1679	0.1062	0.0600	3169.0	0.0027
1.75	20	88	460	4069	0.5092	0.3369	0.1747	0.0936	0.0565	3435.5	0.0036
2.00	20	89	314	4586	0.5118	0.3524	0.1670	0.0950	0.0609	3642.0	0.0049
1.00	25	95	1854	2747	0.4816	0.3492	0.1602	0.1156	0.0582	3883.0	0.0016
1.25	25	100	1074	3454	0.4939	0.3353	0.1530	0.1184	0.0645	3604.5	0.0020
1.50	25	102	800	4174	0.5026	0.3308	0.1540	0.1115	0.0686	3849.5	0.0023
1.75	25	104	539	4920	0.5106	0.3229	0.1684	0.0999	0.0685	4109.0	0.0030
2.00	25	100	362	5583	0.5182	0.3246	0.1701	0.0954	0.0667	4409.5	0.0042
1.00	30	101	2192	3220	0.4904	0.3097	0.1659	0.1203	0.0665	4657.0	0.0013
1.25	30	111	1260	4047	0.4977	0.3133	0.1708	0.1087	0.0665	4210.0	0.0017
1.50	30	111	955	4895	0.5046	0.3186	0.1688	0.0980	0.0745	4590.0	0.0019
1.75	30	110	641	5764	0.5095	0.3201	0.1709	0.1042	0.0678	4858.0	0.0026
2.00	30	109	432	6545	0.5141	0.3337	0.1664	0.1004	0.0627	5195.5	0.0036

4.3 Bradford's Law

We began the analysis of Bradford's law by calculating the area under the Bradford's curve, denoted as $Area_B$, using the following formula:

$$Area_B = \frac{1}{2}[(G(r_2)+G(r_1))(\log r_2 - \log r_1) + (G(r_3)+G(r_2))(\log r_3 - \log r_2) + \ldots + (G(r_m)+G(r_{m-1})(\log r_m - \log r_{m-1})] \quad (4.3)$$

Constant α

We denote $A_B = Area_B/[G(r_m)\log(r_m)]$, where the denominator is the largest area possible for the curve. Figure 4.7 illustrates the effect of α and N on A_B, and it shows that these curves are mirror image of Figure 4.5. For example, at $\alpha = 0.01$ (an extreme condition), A_B of different N converges at approximately 0.95. Note that $G(r_m)$ is equivalent to N, and $r_m = 3$ when $\alpha = 0.01$. We can demonstrate that this is inherently true using similar method as in Section 4.2. On the other hand, as α increases A_B decreases at approximately the same rate across all N until α reaches another extreme condition of 0.99, then A_B shows sudden increases, yet still at about the same rate for all N. Results of $Area_B$ and A_B are summarized in Table 4.5.

Based on Figure 4.7, we can easily visualize the maximum, minimum, and the point where A_B is approximately 50% of all possible area (in this case, $N\cdot\log(r_m)$). Using N = 20,000 as an example, these points are approximately at $\alpha = 0.01$, 0.90, and 0.20, respectively.

When $\alpha = 0.20$ Bradford's law holds, since at this α, $A_B \approx 50\%$ and the curve is near linear with a positive slope (Figure 4.8). As α decreases from 0.20, two things happen. First, G(1) increases; and second, the curve moves northwesterly and causes

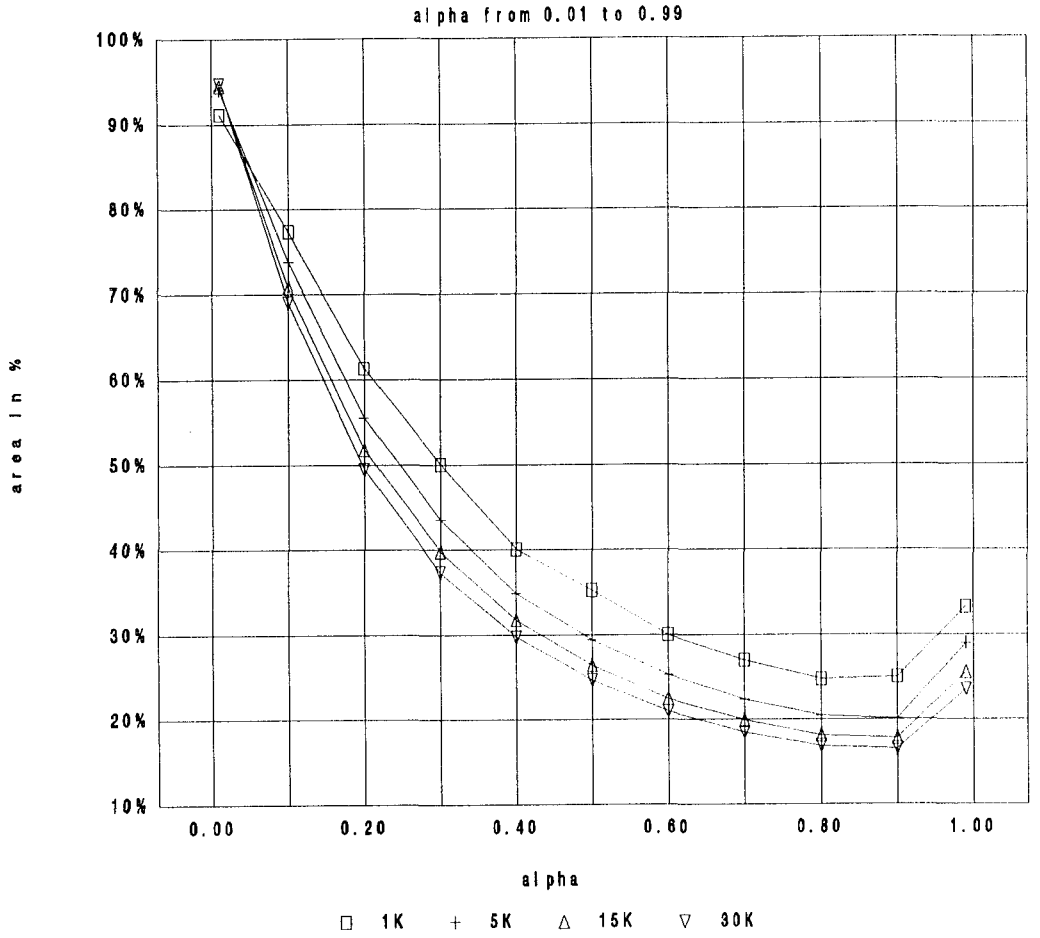


Figure 4.7: Results of Areas under Bradford's Curve (A_B), Constant α

Table 4.5: Simulation Results of Bradford's and Zipf's Law, Constant α

α	N (000)	log(r)	log(g(r))	$Area_B$	$Area_Z$	A_B	A_Z
0.01	1	0.954	2.728	869.9	1.441	0.9118	0.5537
0.10	1	1.944	2.569	1505.0	2.328	0.7742	0.4661
0.20	1	2.262	2.243	1387.5	2.559	0.6134	0.5044
0.30	1	2.480	1.959	1241.8	2.505	0.5007	0.5155
0.40	1	2.603	1.643	1044.3	2.361	0.4012	0.5521
0.50	1	2.695	1.544	951.0	2.182	0.3529	0.5245
0.60	1	2.786	1.204	835.9	1.894	0.3000	0.5646
0.70	1	2.855	1.114	770.6	1.669	0.2699	0.5249
0.80	1	2.903	0.903	719.1	1.163	0.2477	0.4435
0.90	1	2.957	0.699	739.5	0.947	0.2501	0.4581
0.99	1	2.994	0.477	996.6	0.708	0.3329	0.4958
0.01	5	1.643	3.419	7729.3	2.316	0.9409	0.4123
0.10	5	2.676	3.209	9883.8	3.827	0.7387	0.4457
0.20	5	2.996	2.808	8320.8	4.064	0.5555	0.4831
0.30	5	3.183	2.481	6935.9	3.932	0.4358	0.4979
0.40	5	3.304	2.037	5766.2	3.677	0.3490	0.5464
0.50	5	3.396	1.826	5001.8	3.348	0.2946	0.5399
0.60	5	3.471	1.415	4394.3	2.937	0.2532	0.5980
0.70	5	3.538	1.279	3960.6	2.551	0.2239	0.5638
0.80	5	3.597	1.079	3682.7	2.128	0.2048	0.5482
0.90	5	3.652	0.845	3661.6	1.674	0.2005	0.5424
0.99	5	3.695	0.477	5352.6	0.824	0.2897	0.4674
0.01	10	2.021	3.719	19072.6	2.934	0.9437	0.3904
0.10	10	2.999	3.479	21552.7	4.604	0.7187	0.4413
0.20	10	3.305	3.049	17572.8	4.827	0.5317	0.4790
0.30	10	3.482	2.704	14307.4	4.654	0.4109	0.4943
0.40	10	3.692	2.241	11826.1	4.343	0.3203	0.5249
0.50	10	3.692	1.973	10172.9	3.946	0.2755	0.5417
0.60	10	3.773	1.556	8855.8	3.425	0.2347	0.5834
0.70	10	3.841	1.398	7961.5	2.964	0.2073	0.5520
0.80	10	3.900	1.079	7370.7	2.420	0.1890	0.5751
0.90	10	3.955	0.845	7325.3	1.905	0.1852	0.5700
0.99	10	3.996	0.477	10770.8	0.938	0.2695	0.4921
0.01	15	2.173	3.899	30812.9	3.292	0.9453	0.3885
0.10	15	3.177	3.642	33737.3	5.096	0.7079	0.4404
0.20	15	3.483	3.189	27084.5	5.310	0.5184	0.4781
0.30	15	3.657	2.812	21828.4	5.113	0.3979	0.4972
0.40	15	3.773	2.342	17970.1	4.763	0.3175	0.5390
0.50	15	3.869	2.076	15322.2	4.306	0.2640	0.5361
0.60	15	3.950	1.663	13320.8	3.746	0.2248	0.5703
0.70	15	4.020	1.447	11983.8	3.243	0.1987	0.5575
0.80	15	4.078	1.176	11097.0	2.679	0.1814	0.5586
0.90	15	4.130	0.903	10996.1	2.092	0.1775	0.5609
0.99	15	4.172	0.477	16015.0	1.013	0.2559	0.5090

Table 4.5 Continued

α	N (000)	log(r)	log(g(r))	$Area_B$	$Area_Z$	A_B	A_Z
0.01	20	2.312	4.024	43742.4	3.552	0.9460	0.3818
0.10	20	3.295	3.758	46252.9	5.465	0.7019	0.4414
0.20	20	3.601	3.287	36674.3	5.672	0.5092	0.4792
0.30	20	3.780	2.894	29274.2	5.443	0.3872	0.4975
0.40	20	3.904	2.415	24039.3	5.051	0.3079	0.5357
0.50	20	3.996	2.137	20513.4	4.569	0.2567	0.5351
0.60	20	4.076	1.681	17826.6	3.966	0.2187	0.5789
0.70	20	4.145	1.491	16004.0	3.418	0.1931	0.5531
0.80	20	4.202	1.176	14797.3	2.837	0.1761	0.5742
0.90	20	4.256	0.903	14669.8	1.937	0.1723	0.5039
0.99	20	4.297	0.477	21098.5	1.000	0.2455	0.4877
0.01	25	2.391	4.121	56597.4	3.768	0.9468	0.3824
0.10	25	3.386	3.843	58938.7	5.755	0.6963	0.4423
0.20	25	3.698	3.368	46442.2	5.959	0.5023	0.4784
0.30	25	3.877	2.962	36876.2	5.708	0.3805	0.4970
0.40	25	4.000	2.470	30247.3	5.300	0.3025	0.5365
0.50	25	4.094	2.170	25685.6	4.781	0.2510	0.5381
0.60	25	4.173	1.699	22281.4	4.138	0.2136	0.5836
0.70	25	4.242	1.505	19976.8	3.550	0.1884	0.5561
0.80	25	4.300	1.204	18483.7	2.922	0.1719	0.5644
0.90	25	4.352	0.903	18321.9	2.009	0.1684	0.5113
0.99	25	4.394	0.477	26372.8	0.989	0.2401	0.4717
0.01	30	2.483	4.197	70564.8	3.942	0.9473	0.3783
0.10	30	3.465	3.911	71836.7	6.004	0.6911	0.4431
0.20	30	3.777	3.428	56089.2	6.201	0.4950	0.4789
0.30	30	3.956	3.013	44354.2	5.928	0.3737	0.4973
0.40	30	4.079	2.515	36381.9	5.499	0.2973	0.5360
0.50	30	4.173	2.207	30854.8	4.955	0.2465	0.5380
0.60	30	4.253	1.708	26746.7	4.285	0.2096	0.5899
0.70	30	4.322	1.505	23961.0	3.660	0.1848	0.5627
0.80	30	4.380	1.204	22163.9	3.013	0.1687	0.5713
0.90	30	4.432	0.903	21980.6	2.065	0.1653	0.5159
0.99	30	4.473	0.477	31608.6	0.951	0.2356	0.4459

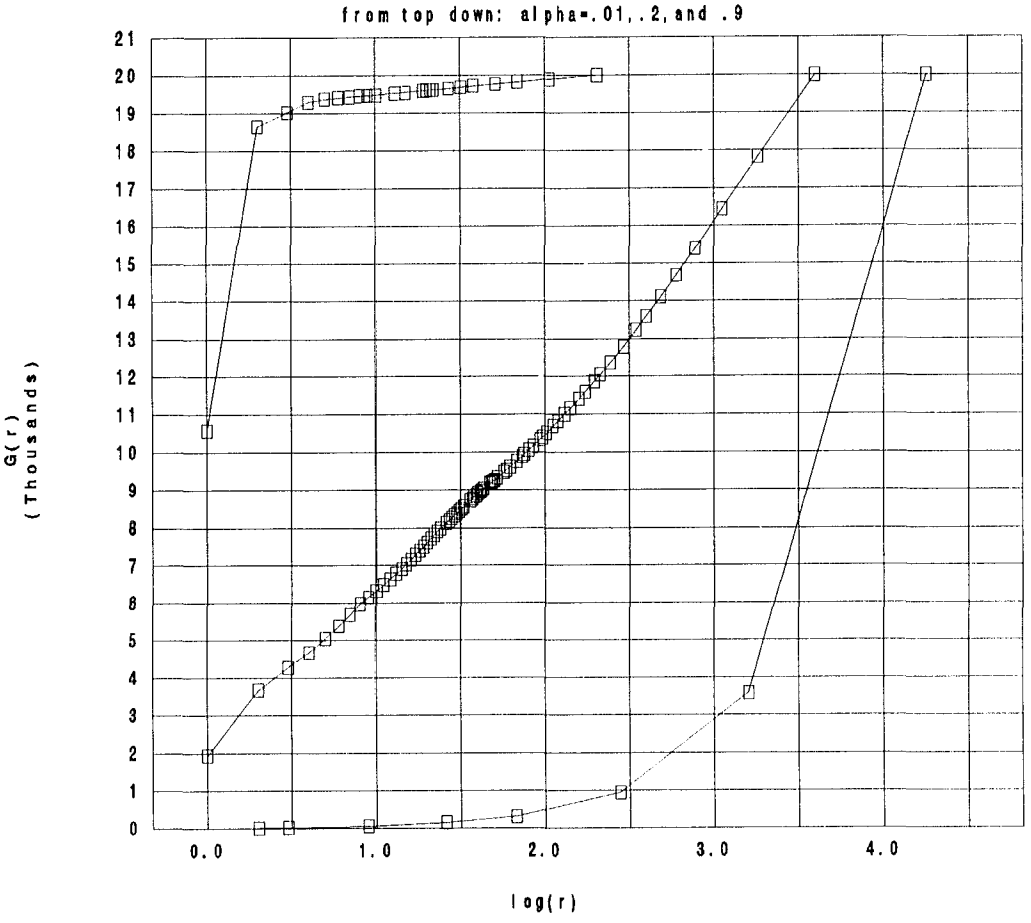


Figure 4.8: Results of Bradford's Law, Constant α

$Area_B$ to increase. Most of the drastic slope changes take place at the *first* few points on the curve, then the curve return to its linear form, though with a flatter slope. The slope of the curve where r_i is large is directly related to α, while the slope of the curve where r_i is small is inversely related to α. Note that the 80/20 rule also holds when $\alpha \approx 0.20$.

In terms of the six different classes of Bradford curves illustrated in Figure 1.1, we are able to reproduce four of the curves using Simon's basic model alone. As expected, the 4th class, with its linearity, can be reproduced using $\alpha = 0.18$. The first class requires $0.2 < \alpha < 0.5$, the 3rd class uses $\alpha > 0.5$, and the 6th class is approximated when $\alpha = 0.10$.

Decreasing α

Table 4.6 summarizes the values of $Area_B$ and A_B at different levels of A and N, and it shows that A_B decreases as A (thus alpha) increases, independent of N. In fact, we selected N = 1,000, 20,000, and 30,000 for illustration purpose, and the three curves basically overlap each other. Since the "all possible area" can also be expressed as $N \cdot \log(r_m)$, it automatically increases when the total number of usage N increases; however, $Area_B$ (the nominal area) changes proportional to N, thus A_B appears to remain unaffected by changing N.

Based on Figure 4.9, the maximum, minimum, and the 50% points of A_B are determined to be at approximately A = 1.0, 2.0, and 1.25, respectively. Figure 4.9 is the composite graph of these three curves. Similar to Figure 4.8, the 50% curve (A = 1.25) is the most linear one, the minimal curve (A = 2.00) bends southeasterly, and the

Table 4.6: Simulation Results of Bradford's and Zipf's Law; Decreasing α, α = A/ln(R) where R=1,2,...,N

A	N(000)	log(r)	log(g(r))	$Area_B$	$Area_Z$	A_B	A_Z
1.00	1	2.228	2.004	1237.2	2.615	0.5553	0.5856
1.25	1	2.326	1.792	1138.4	2.575	0.4894	0.6177
1.50	1	2.433	1.708	1107.0	2.510	0.4550	0.6039
1.75	1	2.501	1.591	1020.4	2.406	0.4080	0.6046
2.00	1	2.542	1.447	963.0	2.323	0.3788	0.6317
1.00	5	2.836	2.629	7908.0	4.153	0.5577	0.5570
1.25	5	2.927	2.405	7233.0	4.119	0.4942	0.5851
1.50	5	3.014	2.288	6777.1	4.052	0.4497	0.5875
1.75	5	3.086	2.124	6304.0	3.939	0.4086	0.6010
2.00	5	3.136	1.987	5949.7	3.830	0.3794	0.6147
1.00	10	3.101	2.900	17267.9	4.943	0.5568	0.5497
1.25	10	3.198	2.678	15798.8	4.912	0.4940	0.5735
1.50	10	3.276	2.562	14699.3	4.840	0.4487	0.5767
1.75	10	3.349	2.401	13693.6	4.721	0.4089	0.5871
2.00	10	3.403	2.260	12921.3	4.600	0.3797	0.5981
1.00	15	3.255	3.058	27166.2	5.444	0.5564	0.5469
1.25	15	3.348	2.829	24847.5	5.415	0.4948	0.5717
1.50	15	3.431	2.712	23168.7	5.342	0.4502	0.5741
1.75	15	3.505	2.551	21513.1	5.213	0.4092	0.5830
2.00	15	3.556	2.394	20231.1	5.087	0.3793	0.5976
1.00	20	3.358	3.176	37419.1	5.816	0.5572	0.5453
1.25	20	3.457	2.944	34234.8	5.786	0.4952	0.5686
1.50	20	3.537	2.828	31856.0	5.712	0.4503	0.5710
1.75	20	3.609	2.663	29544.1	5.584	0.4093	0.5810
2.00	20	3.661	2.497	27680.7	5.448	0.3780	0.5960
1.00	25	3.439	3.268	47879.2	6.113	0.5569	0.5439
1.25	25	3.538	3.031	43844.0	6.085	0.4957	0.5674
1.50	25	3.621	2.903	40739.6	6.011	0.4500	0.5718
1.75	25	3.692	2.732	37749.6	5.878	0.4090	0.5828
2.00	25	3.747	2.559	35432.5	5.743	0.3782	0.5989
1.00	30	3.508	3.341	58540.8	6.364	0.5563	0.5430
1.25	30	3.607	3.100	53549.2	6.333	0.4949	0.5664
1.50	30	3.690	2.980	49694.1	6.257	0.4489	0.5690
1.75	30	3.761	2.807	46031.2	6.122	0.4080	0.5799
2.00	30	3.816	2.635	43142.1	5.979	0.3769	0.5947

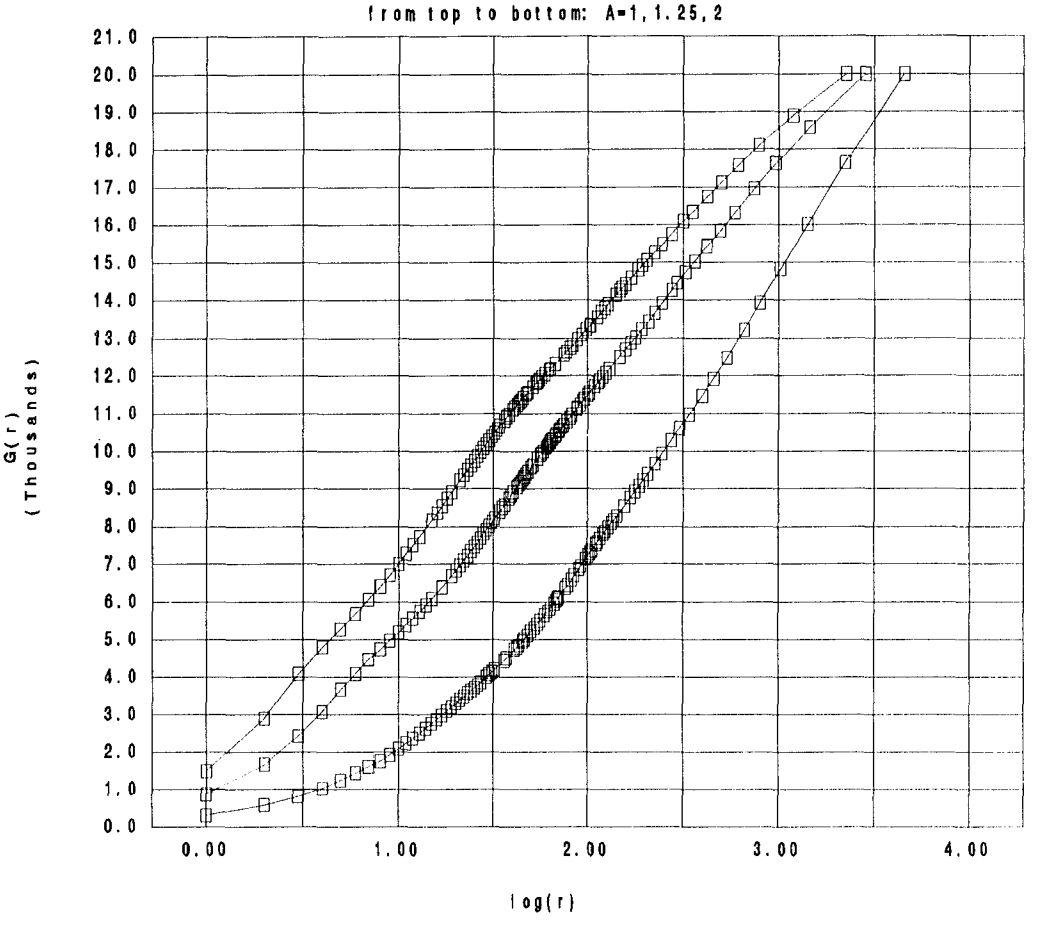


Figure 4.9: Results for Bradford's Law, α = A/ln(R), R = 1,2,...,20000

maximal curve bends northwesterly. Again, when α increases both $Area_B$ and A_B decrease, and vice versa.

4.4 Zipf's Law

The Area under Zipf's curve ($Area_Z$) is calculated as follows:

$$Area_Z = \frac{1}{2}[(\log g(r_2) + \log g(r_1))(\log r_2 - \log r_1) + (\log g(r_3) + \log g(r_2))(\log r_3 - \log r_2) + \ldots + (\log g(r_m) + \log g(r_{m-1}))(\log r_m - \log r_{m-1})] \quad (4.4)$$

We also define $A_Z = Area_Z/[(\log(g(r_m))(\log(r_m))]$, where the denominator is the largest possible $Area_Z$.

Constant α

Figure 4.10 shows how A_Z varies under different α and N. The pattern here is less clearly defined, especially when N = 1,000; however, the changes in A_Z with respect to α for all N still follow a general pattern. The effect of N on A_Z also shows greater dispersions than those described in previous sections. The values of $Area_Z$ and A_Z are summarized in Table 4.6. The nominal values of the area, $Area_Z$, follows the general pattern of A_Z: as α increases, it increases also; however, after reaching certain maximum point it eventually decreases.

Based on Figure 4.10, again we select the minimum, the maximum, and the 50% points of A_Z at α = 0.01, 0.60, 0.30, respectively. We also add to our graph two other important points: the other extreme point α = 0.99, and α = 0.2 where the graph is near linear and has a slope $\approx$ -1. These Zipf's curves are plotted in the same graph in Figure 4.11. In this case, because of the "kink" at the very beginning of the curve, the point A_Z = 50% does not correspond to the ideal condition in which Zipf's law holds.

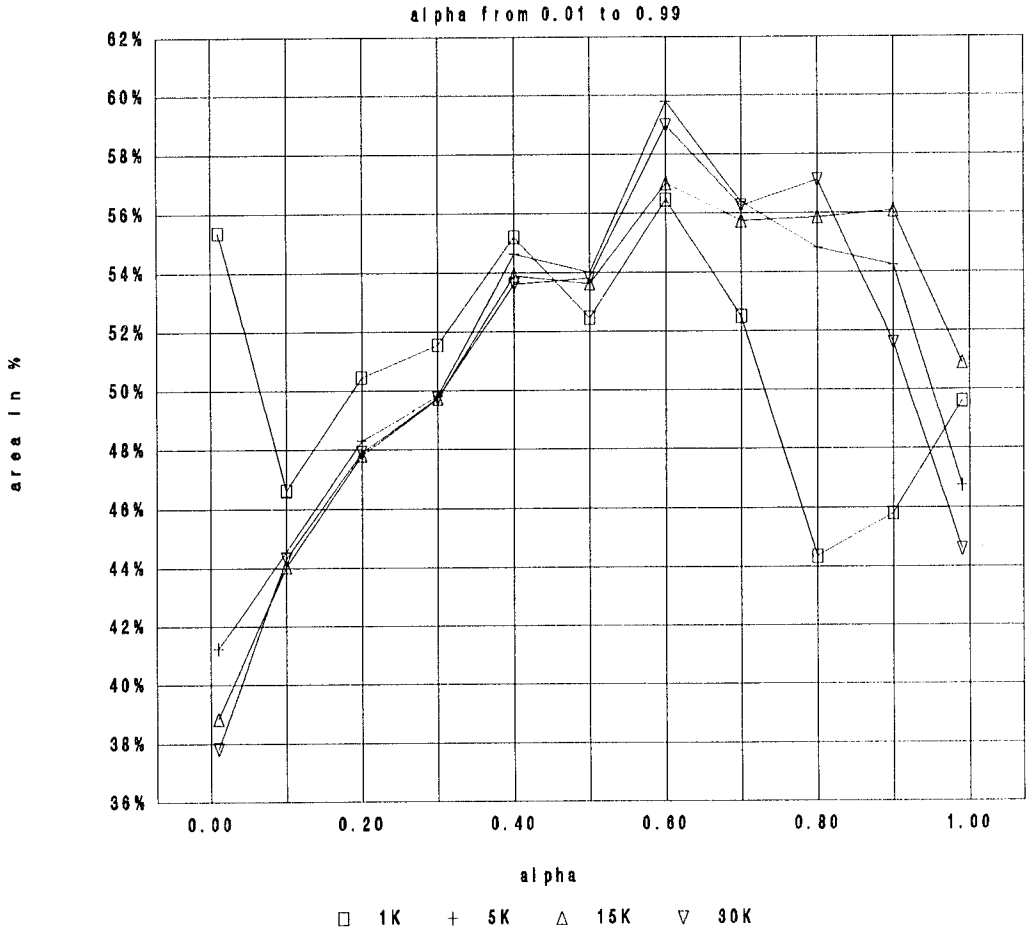


Figure 4.10: Area under Zipf's Law (A_Z)

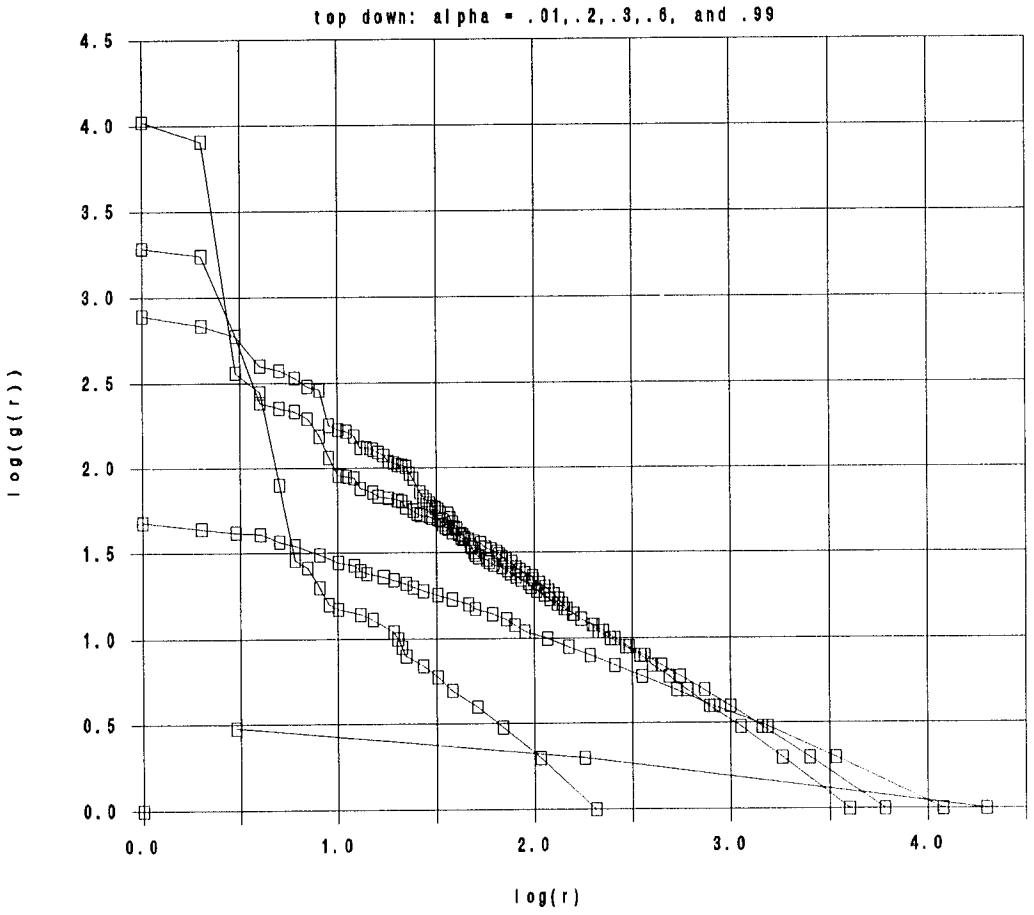


Figure 4.11: Results for Zipf's Law, Constant α

Figure 4.11 can be analyzed from several angles. First, the general negative slope remains to be the characteristics of these Zipf's curves; however, the slope flattens with the increase of α. Second, the initial "kink" in the Zipf's curve remains but becomes less pronounced as α increases. Third, $\log(g(r_i))$ decreases when α increases, but $\log(r)$ increases when α increases; thus, the "largest area possible" as we have used previously changes its shape from an vertical rectangles to horizontal rectangles. Note that similar to the results obtained from previous sections, when $\alpha = 0.60$ and $A_Z \approx 50\%$, the curve is near linear. However, the curve is also near linear when $\alpha = 0.99$, partly due to the much fewer clusters of items with the same usage (m, the maximum index, is 3). On the other hand, coinciding with the 80/20 rule and Bradford's law, Zipf's law holds at $\alpha \approx 0.20$ when the slope approximates -1.

As we have discussed in Section 1.1.4, Zipf's second law expresses the ratios among the number of items of those with low usages, and these ratios can be reproduced in our simulations. Table 4.3 shows that when $\alpha = 0.20$ the ratios ($f(1)/T \approx 0.54$, $f(2)/f(1) \approx 0.43$, $f(3)/f(1) \approx 0.16$, $f(4)/f(1) \approx 0.08$, and $f(5)/f(1) \approx 0.05$) approximate those in Section 1.1.4 (0.5, 0.33, 0.17, 0.10, and 0.07, respectively).

Decreasing α

Table 4.6 is the summary of values of $Area_Z$ and A_Z under different A and N. Note that $Area_Z$ and A_Z are inversely related. Since the minimal $Area_Z$ is greater than 50%, only the minimum and maximum points are selected to plot the Zipf's curves in Figure 4.12.

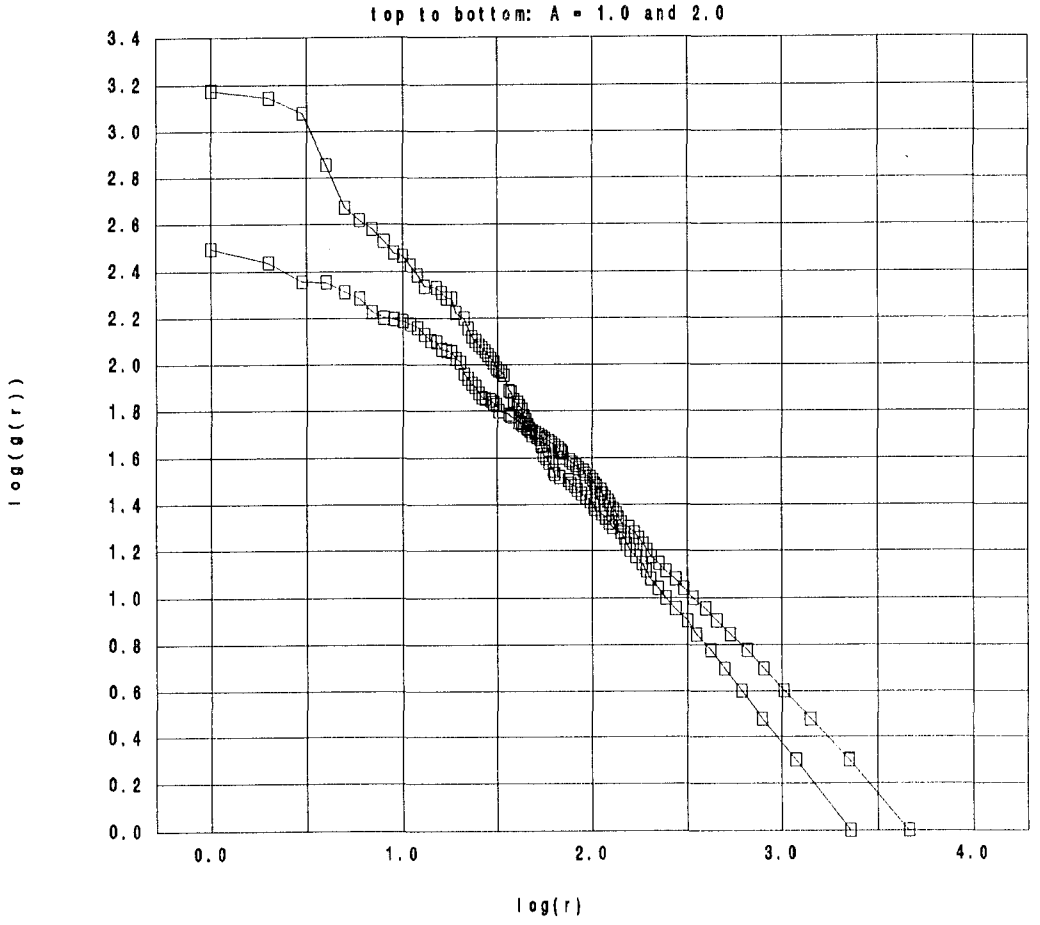


Figure 4.12: Results for Zipf's Law, $\alpha(R) = A/\ln(R)$, R = 1,2,...,20000

When these decreasing functions are evaluated at A = 1.0 and 2.0, and N = 20,000, the minimum values of α are $\approx$ 0.1 and $\approx$ 0.2, respectively, at the end of the iteration. thus, it is no surprise that these two curves lie somewhere around the curve of the constant α = 0.30 in Figure 4.11. Most of the observations made in constant-α model still hold: the flatten slope, the reduction of the initial "kink," the decreasing $\log(g(r_i))$ and increasing $\log(r_i)$ — as α increases.

Table 4.4 shows ratios of low-frequency usage comply very well with the theoretical figures of Zipf's second law. For instance, when A = 1.5 and N = 20,000, the ratios are approximately 0.50, 0.33, 0.17, 0.11, and 0.06 (compare with 0.5, 0.33, 0.17, 0.10 and 0.07 in Section 1.1.4).

4.5 Additional Observations

4.5.1 Verification of the Relationship Between s_m and α

Section 4.1 has shown that the area under the 80/20 curve is inversely related to α, and Figure 4.1 is the visual presentation of this relationship. Table 4.7 shows that the slope between points (x_{m-1},θ_{m-1}) and (100%,100%), or s_m, seem to decrease as α increases. In fact, regression analysis yields $s_m = 0.0001553 + 0.996082\alpha$, with $R^2 = 0.9983$. We thus verify the findings in Section 3.2, and we can estimate the α of an empirical data set through examining its s_m.

Using Kendall's data as an example, we find that $x_{m-1} = 0.451$ and $\theta_{m-1} = 0.885$; therefore, the s_m is determined to be 0.209. We ran a simulation using α = 0.209 and N = 1763 (the total number of papers in Kendall's data), and the result is very similar to the actual data. A better fit is found at α = 0.2225. Table 4.8 compares these three

Table 4.7: s_m at different N and α

N(000)	α	x_{m-1}	θ_{m-1}	s_m
1	0.01	0.556	0.996	0.009
	0.10	0.466	0.953	0.088
	0.20	0.508	0.910	0.183
	0.30	0.401	0.819	0.302
	0.40	0.394	0.757	0.401
	0.50	0.357	0.681	0.496
	0.60	0.280	0.560	0.611
	0.70	0.223	0.444	0.716
	0.80	0.174	0.340	0.799
	0.90	0.079	0.166	0.906
	0.99	0.012	0.025	0.987
10	0.01	0.486	0.995	0.010
	0.10	0.459	0.946	0.100
	0.20	0.452	0.890	0.201
	0.30	0.417	0.823	0.304
	0.40	0.379	0.753	0.398
	0.50	0.340	0.676	0.491
	0.60	0.289	0.578	0.594
	0.70	0.235	0.470	0.693
	0.80	0.171	0.342	0.794
	0.90	0.089	0.179	0.901
	0.99	0.008	0.016	0.992
15	0.01	0.530	0.995	0.011
	0.10	0.474	0.947	0.101
	0.20	0.447	0.888	0.203
	0.30	0.416	0.823	0.303
	0.40	0.385	0.757	0.395
	0.50	0.345	0.678	0.492
	0.60	0.290	0.578	0.594
	0.70	0.228	0.460	0.699
	0.80	0.167	0.336	0.797
	0.90	0.089	0.181	0.899
	0.99	0.009	0.017	0.992
20	0.01	0.522	0.995	0.010
	0.10	0.480	0.949	0.098
	0.20	0.456	0.892	0.199
	0.30	0.418	0.825	0.301
	0.40	0.376	0.700	0.481
	0.50	0.337	0.671	0.496
	0.60	0.287	0.576	0.595
	0.70	0.227	0.461	0.697
	0.80	0.167	0.336	0.797
	0.90	0.089	0.179	0.901
	0.99	0.009	0.018	0.991

$s_m = (1 - \theta_{m-1})/(1 - x_{m-1})$

Table 4.8: Simulated Kendall's Data

	Kendall's		α = .209		α = .2225	
i	n_i	$f(n_i)$	n_i	$f(n_i)$	n_i	$f(n_i)$
1	1	203	1	168	1	198
2	2	54	2	59	2	68
3	3	29	3	30	3	32
4	4	17	4	14	4	16
5	5	10	5	12	5	5
6	6	6	6	4	6	6
7	7	8	7	1	7	7
8	8	8	8	2	8	3
9	9	4	9	6	9	4
10	10	3	10	2	10	1
11	11	5	11	1	11	5
12	12	2	12	3	12	3
13	14	1	13	2	13	1
14	15	2	14	3	14	2
15	16	4	15	2	15	1
16	18	1	16	1	18	3
17	20	2	17	1	20	1
18	21	2	18	2	22	1
19	22	2	23	1	23	1
20	34	1	24	1	25	1
21	49	1	28	1	35	1
22	58	1	35	1	40	1
23	95	1	37	1	41	1
24	102	1	38	1	44	1
25	114	1	46	1	54	1
26	242	1	51	1	101	1
27			54	1	213	1
28			102	1	270	1
29			219	1		
30			279	1		
		----		----		----
T		370		325		367
μ		4.76		5.43		4.80

sets of data. We can see that when $\alpha = 0.2225$, the total number of items T, the total number of index m, and the usage distributions are similar to those in Kendall's data.

If we denote the slope between (0,0) and (x_1,θ_1) to be s_1, then we find that the log of s_1 decreases linearly with α, though not unitarily. Using only N = 20,000 as an example, we have $\log(s_1) = 3.061187 - 2.51188\alpha$, with $R_2 = 0.9946$. However, this measure has little predictive value in estimating α of an empirical data, because, as we will discuss in Chapter 5, it may be affected by other factors also.

4.5.2 Effect of α on m

Figure 4.13 indicates that regardless the level of N, simulation generates the maximum number of indexes (total number of types) with constant $\alpha = 0.20$. Within the scope of our simulation, all indexes are reduced to be three when $\alpha = 0.99$, regardless of the level of N. The maximum number of indexes m increases between $\alpha = 0.01$ to 0.20 then decline at an *increasing* rate as α continue to increase.

4.5.3 Effect of α on n_m

Figure 4.14 shows that n_m declines continuously at a *decreasing* rate as α declines. Similarly, it shows that the highest number of occurrences, n_m, decreases rapidly as α increases. In this instance, at the levels where $\alpha > 0.50$, n_m becomes indistinguishable for all levels of N; in fact, Table 4.3 shows that even with 30,000 usages, no one item is used more than 51 times at $\alpha = 0.60$. We will discuss further the effect of these extreme points to the usage concentration in Chapter 5.

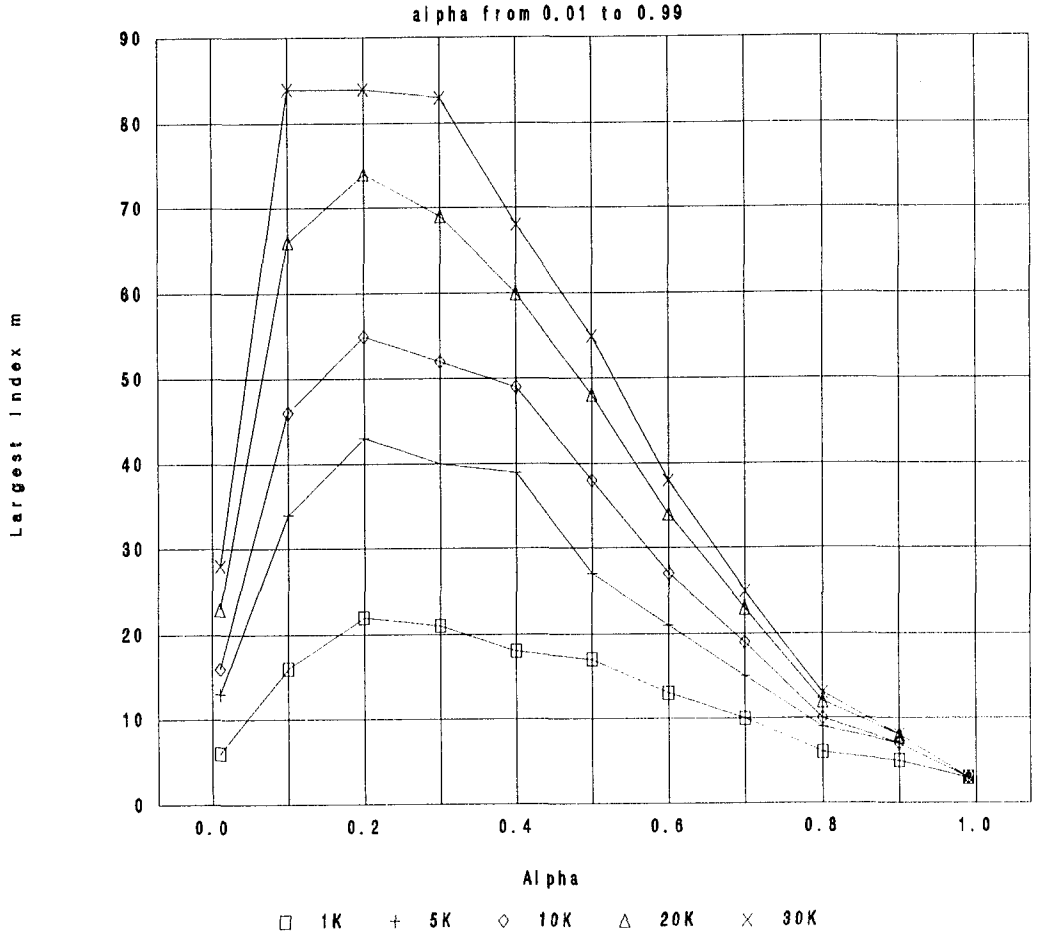


Figure 4.13: The Relationship Between m and α

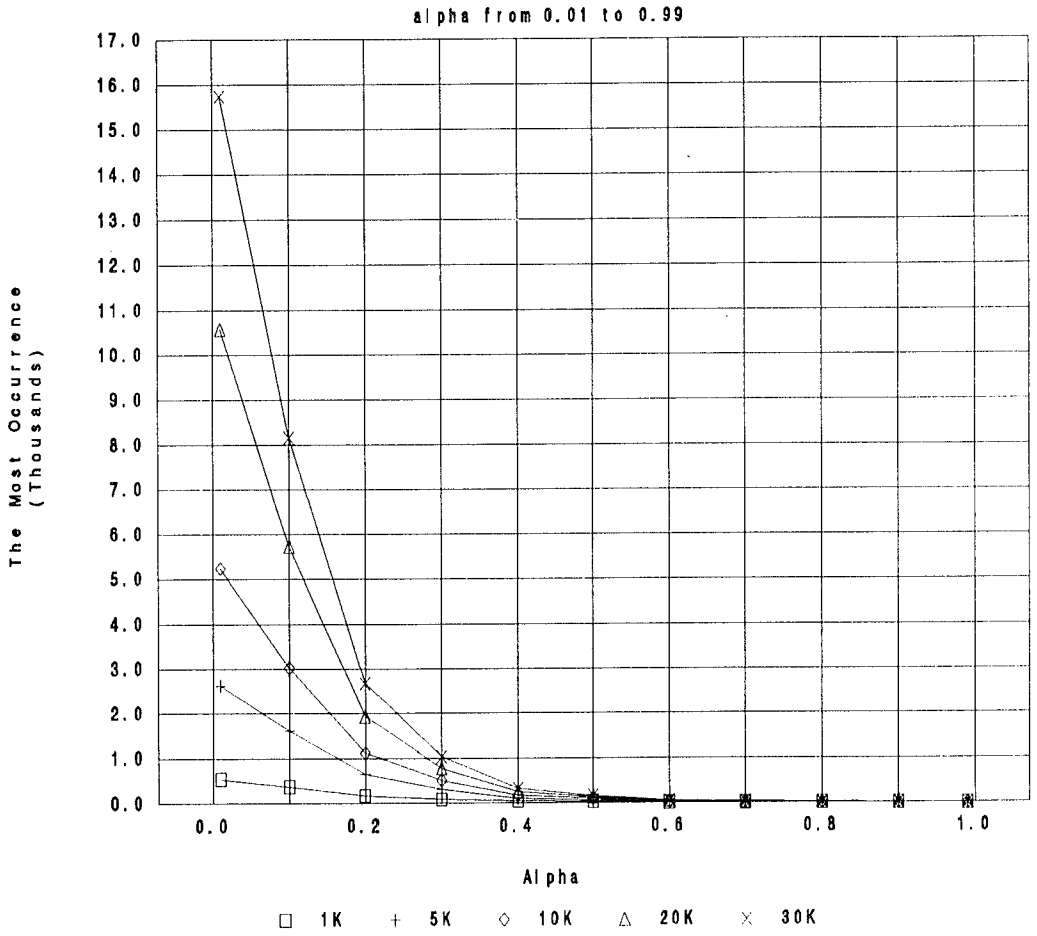


Figure 4.14: The Relationship Between n_m and α

4.5.4 Empirical Phenomena at $\alpha = 0.20$

When $\alpha \approx 0.20$, many things seem to come together. As we have demonstrated in Sections 4.1 to 4.4, the 80/20 rule, Bradford's law, and Zipf's law all hold at $\alpha \approx 0.20$, and Lotka's law holds when $\alpha \approx 0.30$. This reflects the reality that these empirical phenomena are different perspective to the *same* dataset. Also interestingly, the total number of index m also peak at $\alpha = 0.20$, although at this time we do not have an explanation.

The fact that Lotka's law holds at a different α does not necessarily represent a deviation from other empirical phenomena. As we have pointed out at Section 2.1.1, Lotka's law basically focuses on cluster 1, or where many items are used only a few times. In the modified Lotka's curve (e.g., Figure 4.6a), cluster 1 is reflected in the initial linear portion of the curve and where we estimated the curve's slope. Given that cluster 3 disappears with the increase in α, the difficulty in estimating slopes may have contributed to this deviation.

4.6 Summary of Findings

The usage concentration and the shapes of the curves in all four empirical distributions are affected the most by α. We have postulated that since the total number of different items used (T) is the product of α and N, higher α increases the number of items to distribute the usage, therefore decreases the concentration. On the other hand, higher α decreases n_m, the usage frequency of the most active item.

If we hold α unchanged, the time span does not affect the usage pattern. However, in the cases of decreasing function, larger N causes the α to become smaller eventually, and the distributions are affected accordingly.

The simulation results in Sections 4.1 to 4.4 show that these empirical laws hold only within a very narrow range of α values. Specifically, 80/20 rule, Bradford's law, and Zipf's law hold when α is around 0.20; and Lotka's law holds when $\alpha \approx 0.30$. On the one hand, this narrowness suggests the close relationship among these laws; on the other hand, when the probability of new entry deviates from this narrow range then the appropriateness of assuming these laws hold need to be reexamined. While it is possible that many data satisfy the restriction of having $\alpha \approx 0.20$, researches based on Simon's model is more robust because it does not require these restrictive assumptions on data distributions at all. As we have also demonstrated in Section 4.3, we can generate usage patterns of four of the six known Bradford's curve by changing α alone. In fact, these four curves represent a near complete spectrum of Bradford's curve represented by Simon's basic models.

Paving the way for Chapter 6 where we demonstrate our findings using Simon's model to library weeding policy, we verify the validity of the method of estimating α from the empirical data proposed in Section 3.2. This technique allows us to classify systems according to an attribute that directly affects their usage patterns.

CHAPTER 5

COMPUTATIONAL RESULTS OF SIMON'S AUTOREGRESSIVE MODEL

Chapter 4 analyzes the simulation results of the 80/20 rule and other empirical phenomena, and we find that a higher probability of new entry α reduces the concentration of information usage. This chapter shows the results of using Simon's autoregressive model to incorporate the "decay" factor γ into the simulation. We explore how γ affects the shape of the curves and usage concentration as demonstrated in the 80/20 rule and other empirical phenomena, and we find that Simon's basic model with constant α represents an upper bound for the concentration measures generated from the autoregressive model. In light of this, we further specify the limiting conditions, with respect to γ, under which these empirical phenomena would hold as formulated.

5.1 The 80/20 Rule

In terms of Area of the 80/20 curves (Equation 4.1), when α and γ are held constant, the total number of usages N does not affect the outcomes beyond a certain point in our simulation (see Figure 5.1). It seems that this stochastic process yields stable results when $N \geq 15{,}000$; thus, we arbitrarily chose the same $N = 20{,}000$ to show our simulation results. Furthermore, since we have determined in Chapter 4 that when $\alpha \approx 0.20$, the 80/20 rule, Bradford's law, and Zipf's law hold; and Lotka's law is approximated, we demonstrate our simulation results centering on holding $\alpha = 0.20$.

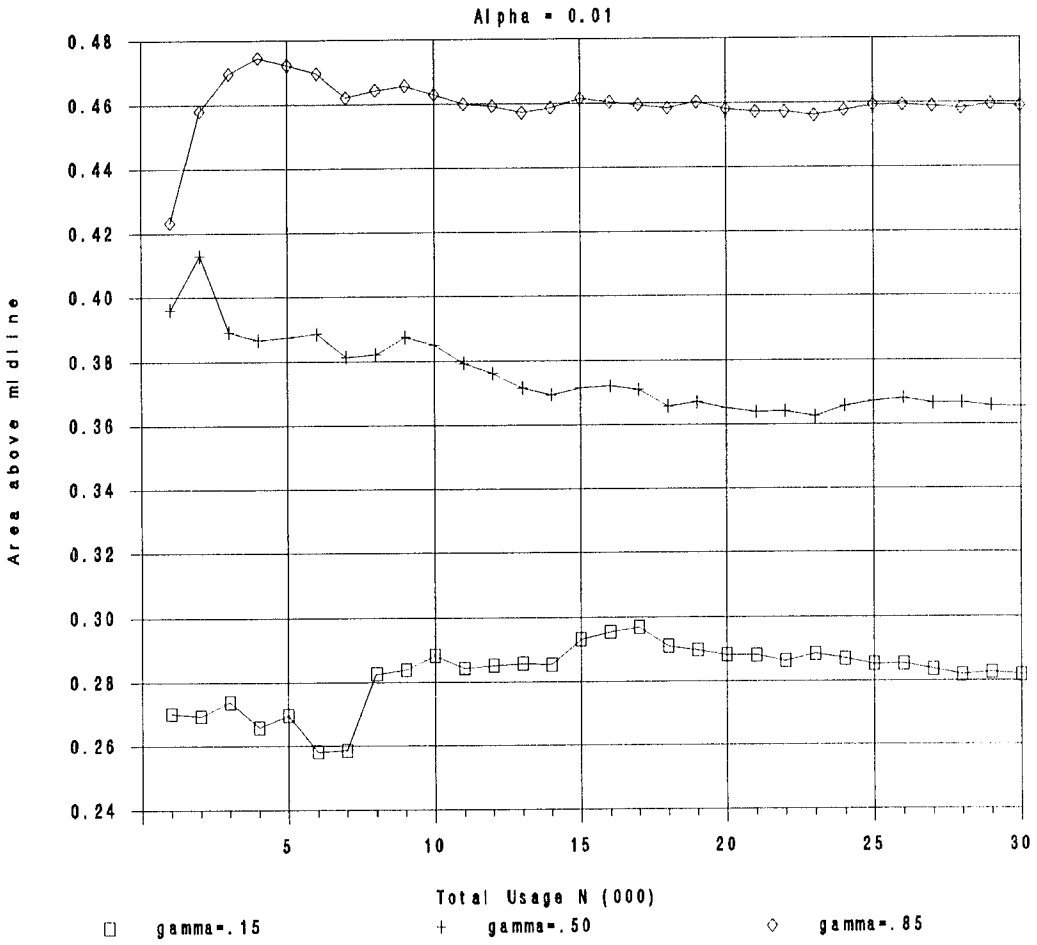


Figure 5.1: 80/20 Area under Different α, γ, and N

Figure 5.2 shows that 80/20 curves overlap at the "end" points despite the wide range of γ. This is verified in Table 5.1 which shows that $s_m \approx 0.2021$ for all γ; therefore, the method proposed in Section 4.6 for estimating α will still work. For the initial point of 80/20 curves, we find that higher γ increases $\log(s_1)$ almost linearly. Note, however, the big difference between the curves of $\gamma = 0.99$ and $\gamma = 1.0$. Excluding the point of $\gamma = 1.0$, regression analysis yields $\log(s_1) = 0.8141 + 0.57\gamma$, with $R^2 = 0.9426$. Examining only the range of $\gamma = 0.99$ and 1, $\log(s_1)$ and γ seems to have a linear relationship also, though with a much steeper increase. Since high α decreases s_1 geometrically (Section 4.5) and high γ increases it, s_1 has no predictive power without holding one factor constant. Furthermore, simulation results show that when $\gamma = 1.0$ the curve is basically identical to that of the basic model, thus, the big change in concentration is indicative the strong effect of γ on the usage pattern, and it allows us to detect the existence of autoregression by seeing the deviation from the curves obtained through the basic model of constant α.

Consider the meaning of γ: a higher γ means a slower rate of decay; i.e., an item's usage probability is little affected by its not being used. Thus, when $\gamma = 1.0$, there is no decay takes place at all, and the result should be identical to those of basic models. On the other hand, a small γ signifies that items that have not been used recently will possibly be neglected for a long time to come, regardless how active they had been previously. A related interpretation is that previously-inactive items do have a chance to become dominant in the future selection process, albeit perhaps temporarily.

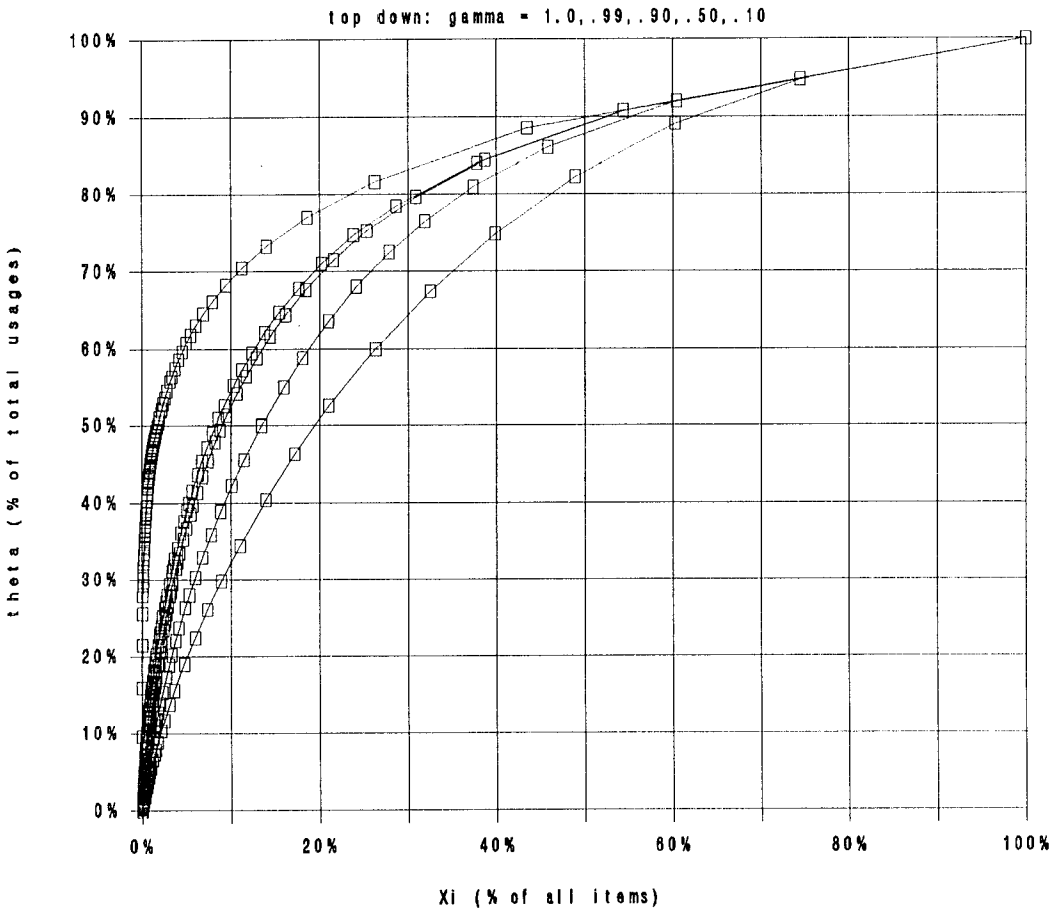


Figure 5.2: Results for 80/20 Rules, $\alpha = 0.20$, $\gamma = 0.1$ to 1.0, N = 20,000

Table 5.1: Slopes of the 80/20 Curves in the Autoregressive Model, $\alpha = 0.2$ and N = 20,000

γ	x_1	θ_1	x_{m-1}	θ_{m-1}	s_m	s_1	$\log(s_1)$
.10	.00025	.00200	.74474	.98840	.2021	8.0	.9031
.20	.00025	.00230	.69330	.93800	.2022	9.2	.9638
.30	.00025	.00235	.65100	.92945	.2021	9.4	.9731
.40	.00025	.00240	.62107	.92341	.2021	9.6	.9823
.50	.00025	.00340	.60425	.92000	.2021	13.6	1.1335
.60	.00025	.00325	.57927	.91495	.2021	13.0	1.1139
.70	.00025	.00390	.56790	.91265	.2022	15.6	1.1931
.80	.00025	.00415	.56369	.91180	.2021	16.6	1.2201
.90	.00025	.00555	.54366	.90775	.2022	22.2	1.3464
.99	.00025	.00690	.54440	.90790	.2022	27.6	1.4409
1.00	.00025	.09585	.43408	.88560	.2021	383.4	2.5837

$s_1 = \theta_1 / x_1$

$s_m = (1 - \theta_{m-1}) / (1 - x_{m-1})$

Therefore, it is plausible that a lower γ (i.e., with high decay rate) should induce less concentration in its usage pattern.

Since $\gamma < 1$ reduces the usage concentration, and under most circumstances it is not plausible that $\gamma > 1.0$ (i.e., the lack of usage of an item increases the probability of its selection); we suggest that Simon's basic model with constant α provides the maximum concentration for a given α. Thus we may conclude that the 80/20 rule is true if and only if $\alpha \leq 0.20$. Since lower γ decreases concentration, our simulation shows that a combination of $\alpha < 0.20$ and some proper γ can also achieve 80/20 (e.g., $\alpha = 0.1$ and $\gamma = 0.8$).

5.2 Lotka's Law

Figures 5.3a through 5.3d show modified Lotka's curves under $\gamma = 0.1$, 0.5, 0.99, and 1.0, respectively. Three observations can be made on the effect of γ. First, notice the lack of region 3 where $\log(f(n_i)) = 0$ (recall that in region 3 $f(n_i) = 1$) in all cases except that of $\gamma = 1.0$ (the basic model); second, the negativity of slope reduces (flattened curves) with the decrease of γ; and third, when $\gamma \leq 0.5$, the linear region 1 becomes curves.

As we have discussed in Section 4.2, the lack of region 3 means few items have extremely high frequency of usage, thus less usage concentration. Recall that at $\alpha \approx 0.30$ this slope is approximately -2 and Lotka's law holds, since a lower γ produces a less steep slope, in order for Lotka's law to be true, it is necessary that $\alpha \leq 0.30$. However, the problem of linearity would arise if $\gamma \leq 0.5$. Therefore, we propose that Lotka's law holds if and only if $\alpha \leq 0.30$ and $\gamma \geq 0.5$.

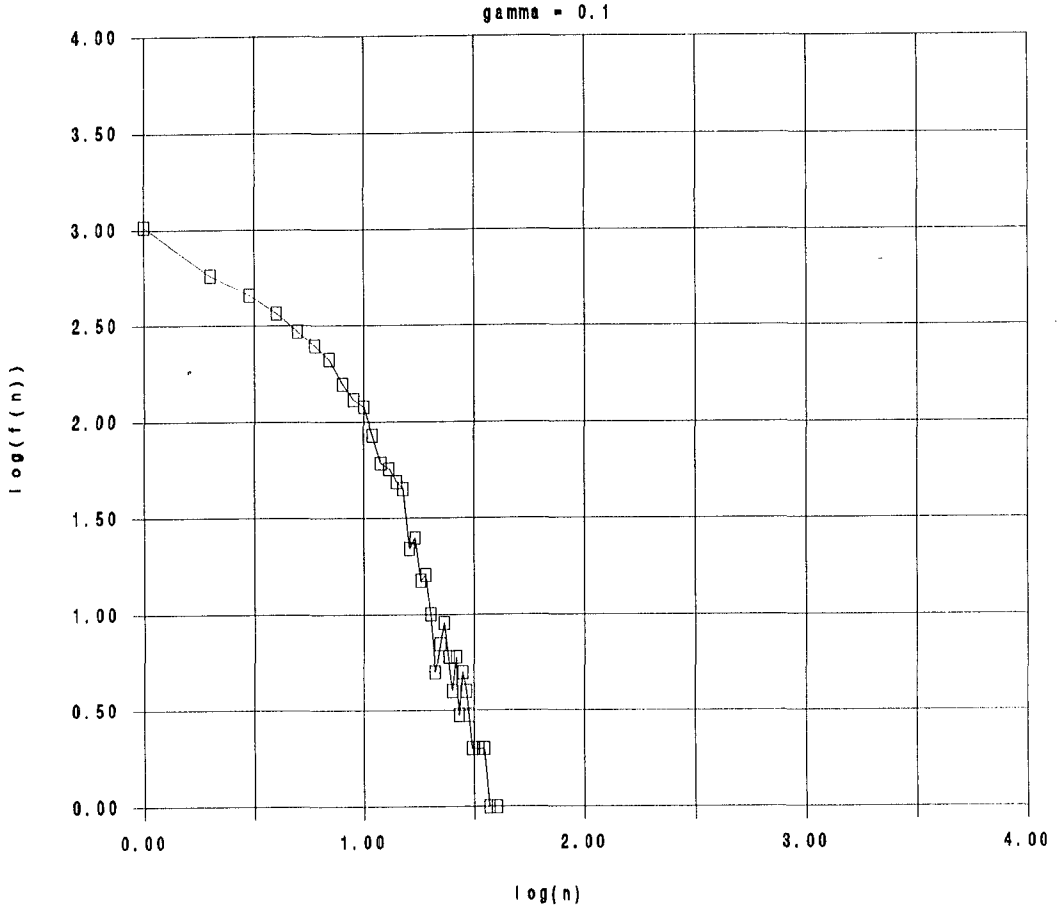


Figure 5.3a: Modified Lotka's Curve, with $\alpha = 0.2$, $\gamma = 0.1$, and N = 20,000

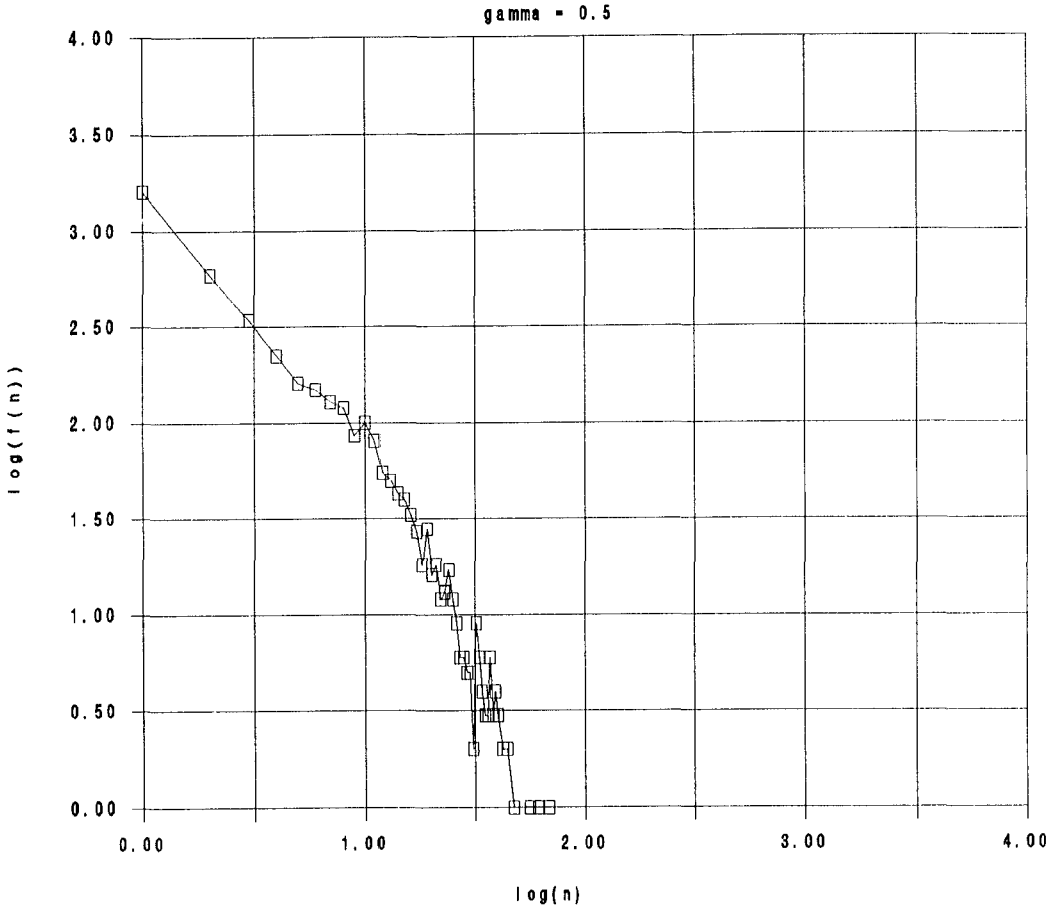


Figure 5.3b: Modified Lotka's Curve, $\alpha = 0.2$, $\gamma = 0.5$, and N = 20,000

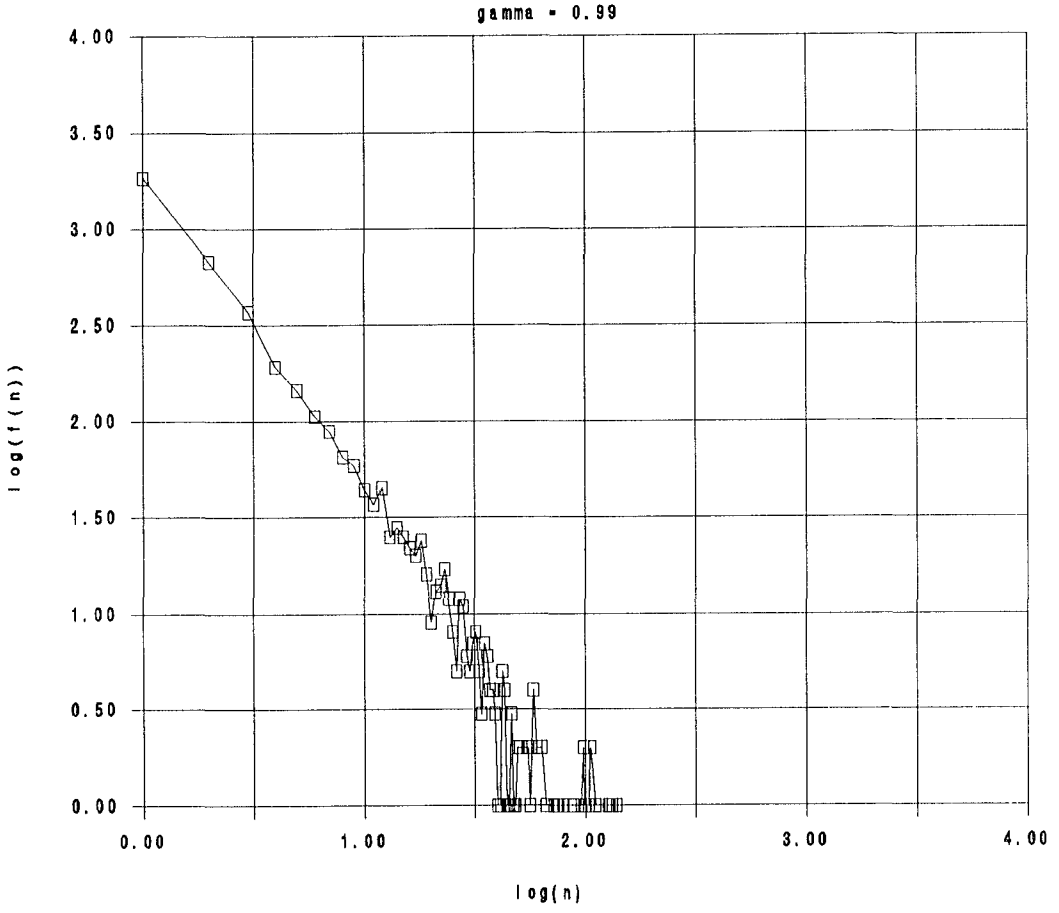


Figure 5.3c: Modified Lotka's Curve, $\alpha = 0.2$, $\gamma = 0.99$, and N = 20,000

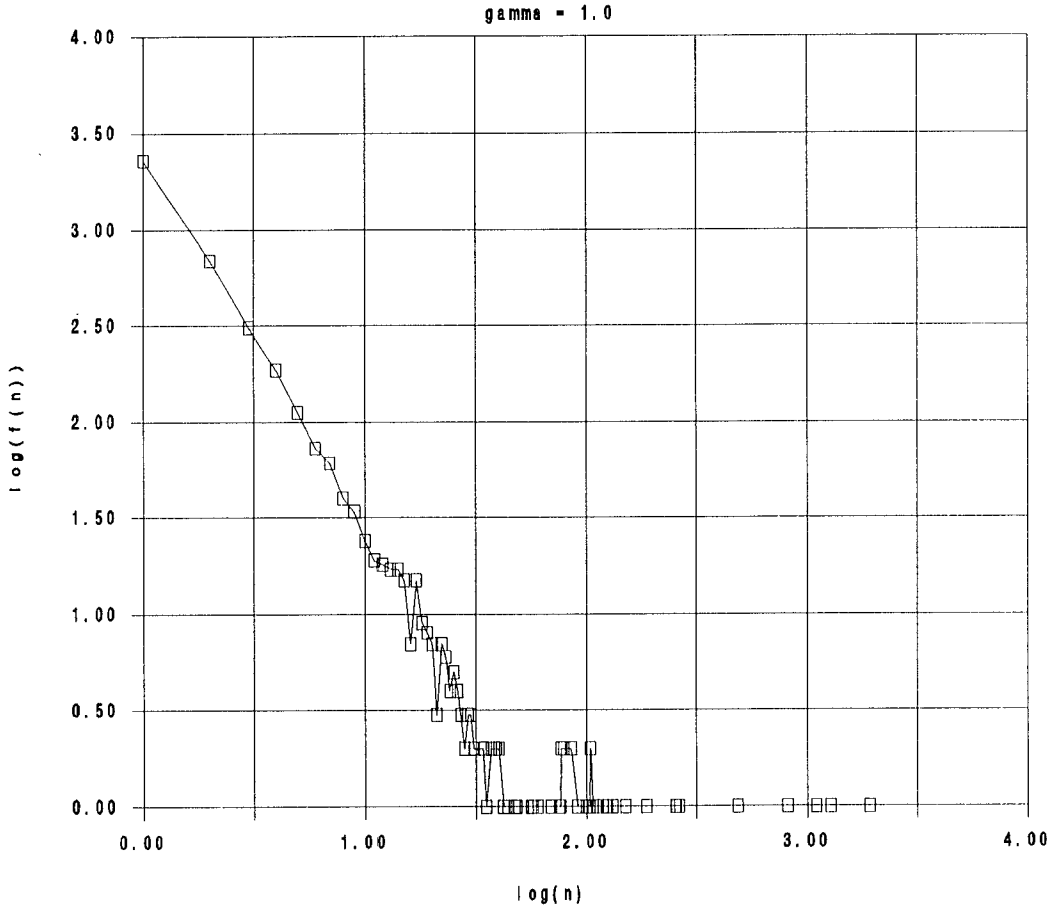


Figure 5.3d: Modified Lotka's Curve, $\alpha = 0.2$, $\gamma = 1.0$, and N = 20,000

5.3 Bradford's Law

Figure 5.4 shows the effect of γ on Bradford's curves. It is amazing that there is basically no resemblance between the basic model (i.e., $\gamma = 1.0$) and any of the curves where $\gamma < 1$. More strikingly, most of the simulation results of $\gamma \leq 0.99$ have a curve that is below the diagonal line that has a slope of 1. In other words, if we use the measurement A_B as defined in Section 4.3, $A_B < 0.5$ when $\gamma \leq 0.99$ for most α. Even in the extreme case of $\alpha = 0.01$ where A_B is large, A_B drops rapidly from 0.945 in the basic model to 0.675 when $\gamma = 0.90$.

Such drastic flattening of the initial points can be seen in Table 5.2. For γ decreases from 1.0 to 0.99, we can compare the first two points on the graph, $G(r_1)$ and $G(r_2)$, and see a steep fall of from 1917 to 138 for $G(r_1)$ and from 3191 to 270 for $G(r_2)$. This drop is reflected in a smaller initial slope and a lower y-intercept for the $\gamma = 0.99$ curve. Since $G(r_1)$ is the usage frequency of the most-used item, and $G(r_2)$ measures the usage of the first two most-used items, this flatten initial slope signals a drastic reduction of usage concentration — a finding that is consistent with our previous analyses. On the other hand, the numbers of seldom-used items f(1) and f(2) (used only once and twice, respectively) also fall. Interestingly, although f(1) declines quickly when γ is low, f(2) does not change much. Since the total number of usage N = 20,000, a combination of decline in f(1) and $G(r_1)$ means less extreme cases of high and low usage, thus creating a usage pattern that has relatively similar amount of items in every usage class.

By adding γ to the simulation we can also produce the second class of Bradford's curve which we could not simulate previously using Simon's basic model alone. The

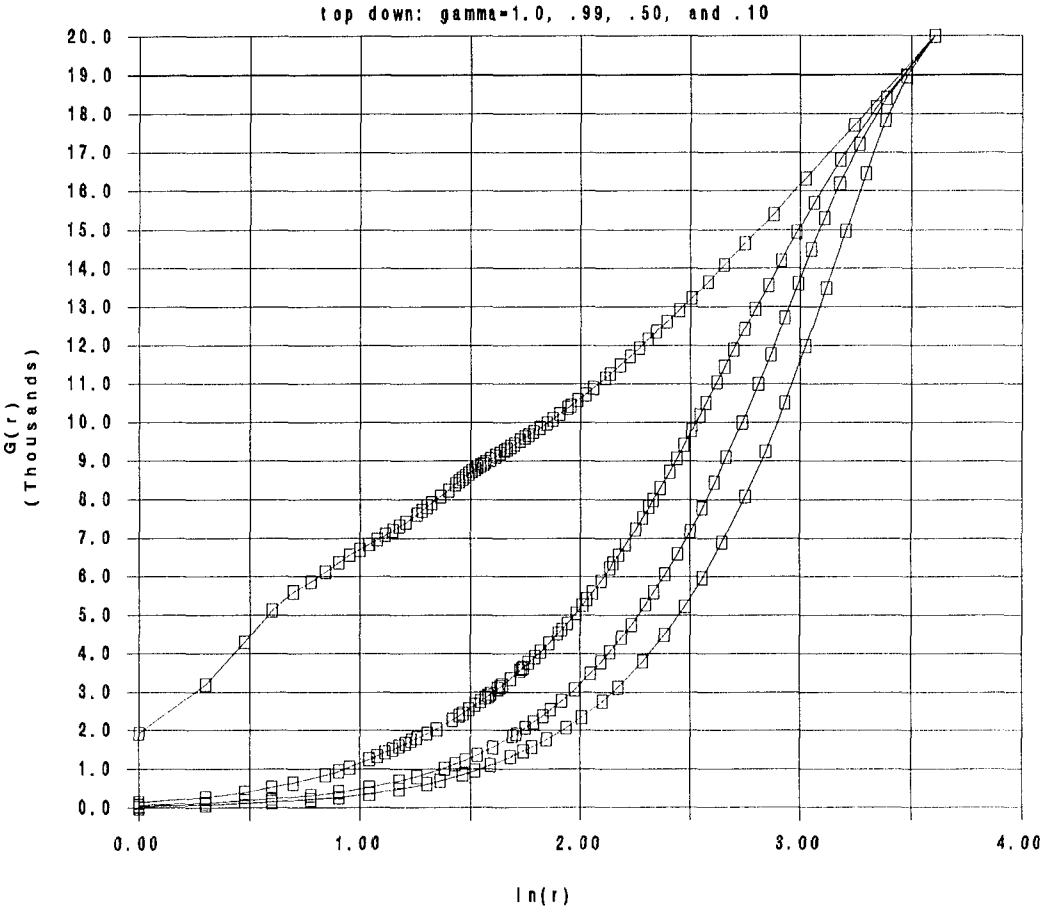


Figure 5.4: Bradford's Curves for $\alpha = 0.2$, $\gamma = 0.1$, 0.5, 0.99, and 1.0

Table 5.2: Initial Points and End Points of Bradford's Curves

γ	m	$G(r_1)$	$G(r_2)$	$G(r_1)/G(r_2)$	f(1)	f(2)
0.10	35	40	77	51.9%	1032	575
0.20	37	46	86	53.5%	1240	567
0.30	41	47	93	50.5%	1411	562
0.40	44	48	95	50.5%	1532	595
0.50	46	68	129	52.7%	1600	591
0.60	50	65	127	51.2%	1701	587
0.70	54	78	220	35.5%	1747	611
0.80	62	83	160	51.9%	1764	664
0.90	71	111	215	51.6%	1845	633
0.95	70	111	208	53.4%	1843	639
0.99	73	138	270	51.1%	1842	674
0.995	74	288	501	57.5%	1839	678
0.9995	74	1213	2007	60.4%	1948	660
1.00	70	1917	3191	60.1%	2288	694

results are presented in Figure 5.5. These curves also highlight the decreasing low-frequency items when γ is low (near the top of the graph). We suspect that the remaining 5th class is a combination of two different Bradford's curves.

5.4 Zipf's Law

Figure 5.6 indicates that similar to the modified Lotka's curves, Zipf's curve flattens and loses its linearity as γ decreases. Recall our finding in Section 4.4 that Zipf's curve has a slope approximates -1 when $\alpha \approx 0.20$, we may conclude that Zipf's law holds as formulated if and only if $\alpha \leq 0.20$. However, as Figure 5.6 indicates, even with $\gamma = 0.99$ the curve begins to lose linearity. Thus, because of the restrictive condition for Zipf's law to hold, it is even more impressive that many data do exhibit Zipf's curve.

As we have indicated in Section 4.4, a flat Zipf's curve means that the usage frequency for an item decreases at a slower rate than its rank. In other words, the infrequently-used words now have higher-than-expected usage relative to Zipf's law as formulated. The lower y-intercept again reflects lower values of $G(r_1)$ (see Table 5.2). On the other hand, when log(g(r)) is small (meaning low usage), we can see the effect of lower f(1) values associated with lower γ in Table 5.2. The steeper slope at the end of Zipf's curve with $\gamma = 0.1$ actually indicates that the difference between f(1) and f(2) is not as much as that of higher γ.

This interpretation reflects our previous findings of γ's effect on usage patterns: fewer items dominate the usage and fewer items are totally neglected when γ is introduced.

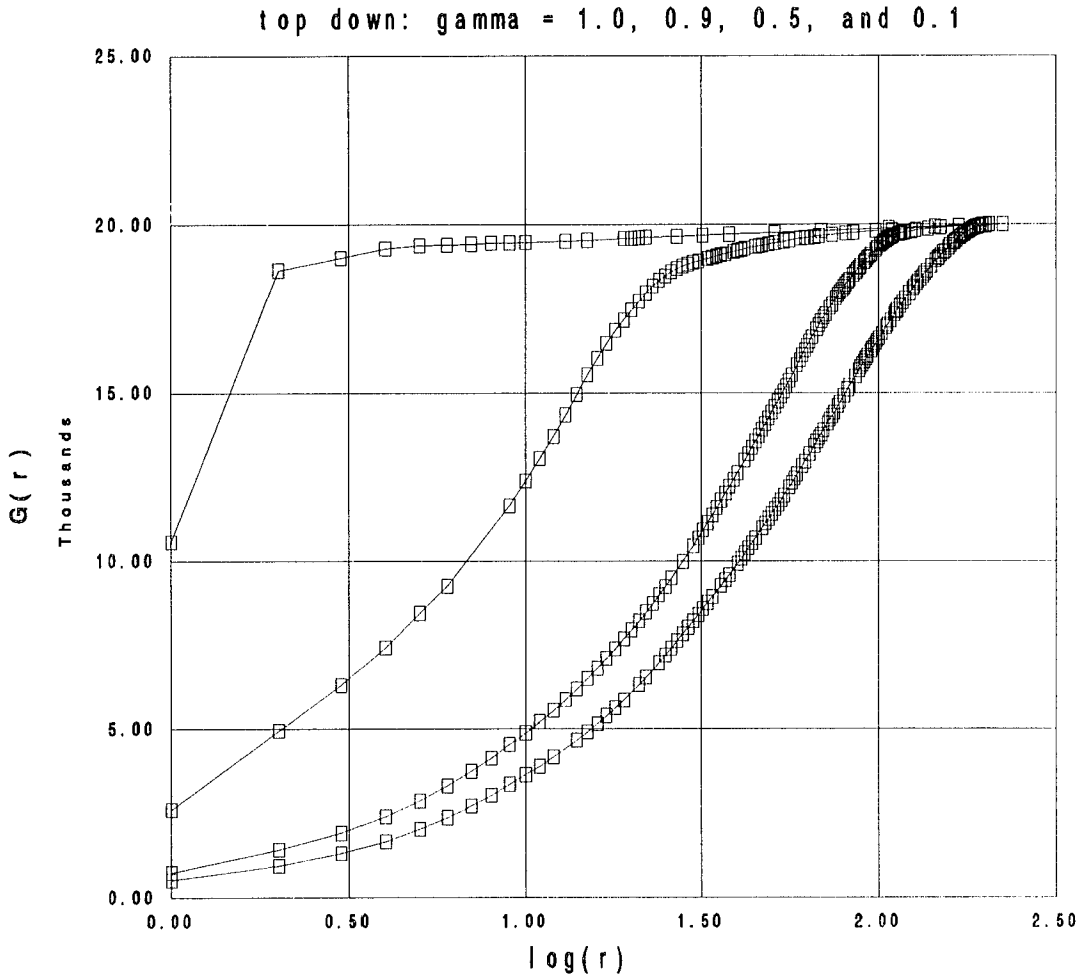


Figure 5.5: Bradford's Curves for $\alpha = 0.01$, $\gamma = 0.1$, 0.5, 0.9, and 1.0

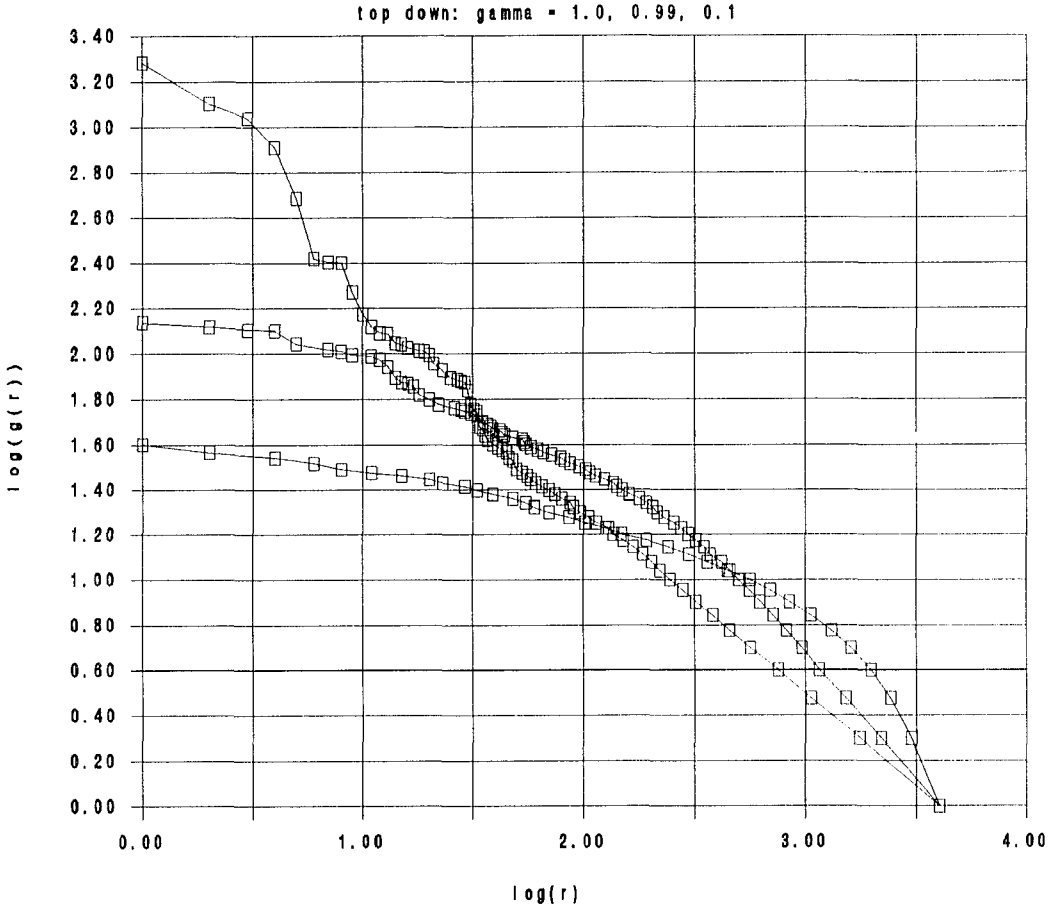


Figure 5.6: Zipf's Curves, α = 0.20, N = 20,000, and γ = 0.1, 0.99 and 1.0

5.5 Summary of Findings

In Simon's two basic models α is the primary force that affects the shapes of the empirical phenomena. Higher α produces lower concentration in usage patterns. We have attributed this phenomenon to the increasing new comers forcing the spreading out of usages. In Simon's autoregressive model, lower γ (higher decay) reduces concentration because it provides an environment in which new comers and the previously low-usage items are not overwhelmed by previously-active items. It is important to note that we hold N and α constant in our model, thus the total number of different items T is fixed in each simulation. The changes in concentration, therefore, solely come from the different distribution of the N usages to these T items.

In the 80/20 curves, changing γ affects the shape of the curves and the usage concentration, but not s_m; thus α can still be estimated independent of γ. Generally, Simon's basic model with constant α provides an upper-bound for the concentration measure, because a positive decay ($\gamma < 1$) reduces the concentration measure.

This reduction of concentration is also evident in the other empirical laws, and it is better shown through analyzing the effect of γ on f(1) and n_m (equivalent to G(1)). It is when any $\gamma < 1$ is introduced into the simulation, these two measures decrease rapidly. This reduction is caused partly because of the total number of iteration N we have selected is quite high. At this number, although the distribution stabilizes, the effect of γ has also been magnified tremendously. However, the implication remains to be that a unused item loses its potential of being used again — no matter how historically active it has been. Thus, not only the frequency of usage is important in determine the

future activity of an item as proposed in Simon's two basic models, the timing of usage is also important in the autoregressive model. Since this usage "decay" or "forgetting" increases the chance for competition from other information items, both f(1) (measuring inactivity or neglect) and n_m (measuring popularity) all fall drastically.

CHAPTER 6

WEEDING LIBRARY COLLECTIONS: AN APPLICATION

In this chapter, we attempt to show an example of applying our findings in this dissertation to weeding process of library collections. Similar process can be applied to other suitable subjects. For example, according to one study, up to 80% of the servers' disks are occupied by inactive files (Stuart 1992). Facing limited disk space, we often have to determine which information to keep active and what to archive. Libraries face similar problems deciding what to keep on shelves and what to keep in storage. This chapter is an application of Simon's model to the weeding process, using the library circulation data at Southeastern Louisiana University as the basis of the study.

6.1 The Need for Weeding Library Collections

Philip Morse (1968) in his landmark book *Library Effectiveness: A Systems Approach* said that the major task of most libraries is to provide the material desired by the majority of its users, and to provide it as quickly and to as many users as is compatible with its budget. On the one hand, the rapidly growing collections, shortage of space, the high cost of storing books on open stacks, and the high costs of new buildings force libraries to consider storage space when acquiring new books (Slote 1989). On the other hand, libraries must remove unneeded items from the shelf to help users find the items of information they need with a minimum of delay and frustration.

This problem of storage space and the cost of search will remain even if all holdings are digitized some day (Morse 1968).

This process of identifying books to discard is called "weeding" (Slote 1989, Evans 1987). Clark (1991) suggests that if a collection contains many items of little interest, those that are useful will not be so readily visible or accessible; therefore, weeding would increase circulation. Furthermore, collections should be weeded to increase the speed of access and to improve the accuracy in retrieval; and finally, those books least likely to be used in the future should be removed to reduce the costs of maintaining a large collection (Slote 1989). If storage space is limited, it would be essential to separate little-used books from a working collection of highly-used ones, and then discarding duplicates, worn out volumes, and obsolete material.

In order to separate "little-used" items from a highly-used "working collection," we need to examine the usage pattern of library materials. In literature, the usage of library resources are found to show the so-called "Matthew Effect" (Merton 1973), i.e., the more a book has been used before, the more likely it will be used again. Burrell (1980) points out that a library holding is selected according to its desirability to users. Environmental factors may cause some subjects to become "hot topics" at a particular time, thus affect the desirability and the usage pattern of the library holdings. The literature also described an "aging" factor in which the probability of usage decays with time (Kent et al. 1979, Burrell 1980, Anderson 1990) perhaps due to obsolescence. Thus, the usage pattern of library holdings follows the empirical phenomena described

in this dissertation that "success breeds success" and is a good candidate for applying Simon's model to describe its characteristics.

6.2 Methods of Weeding

Several weeding criteria are used in libraries: for example, weeding based on appearance or condition of the holdings, upon superfluous or duplicate volumes, or upon age of the holdings alone. The weeding policy varies according to the type of the library and its environment (Kovacs 1990). For example, while weeding and deselection were at a minimum in public libraries because computer data are not widely available, some medical school libraries weed out only duplicated copies. At Southeastern Louisiana University where we collected circulation data, state regulations mandate that books cannot be discarded, therefore holdings are taken out of circulation mostly for their poor appearances.

As we have found out, even with the automated system in place, it is still quite cumbersome for librarians to retrieve information from the database for analysis. A more systematic way of weeding is described below. Specifically, we discuss Slote's (1989) method which has been used extensively in the United States. Although Slote recognized that his method can be fully programmed in an automated system, the computer's role in his example was limited to providing on-line information on the last charged-out date of the holdings in a manual procedure.

Slote's method relies almost exclusively on circulation data to identify possible weeding candidates, while others may take the age of publication into consideration (Evans 1987). However, as Slote (1989) points out, decision making based on the age

of publication is complex and subjective and therefore useless as a systematic approach. To implement Slote's method, we need to acquire the following information from the computer system (Slote 1989, pp. 176-185): (1) The date of previous use, (2) The date of accession the book (i.e., the date the book is added to the system), (3) The number of uses this volume has experienced since the computer circulation control system was initiated, (4) The imprint date of the volume, and (5) A "dusty book" list: all volumes showing no use since the system started up.

Slote's method consists of tabulating the date of the previous use of at least 400 volumes at the circulation desk, representing consecutive volumes being charged out. This sample may be of all classes of books or of any subclasses. He tallies books according to the year the book was last checked out. In the case that the book has not been checked out before, he records the year of the book's accession. He calculates each year's usage as a percentage of the sample and then the cumulative percentage from the current year. A reasonable, arbitrary "keeping level," say 96 percent of sample usage, is established, and a cut-off point is created from these data. Using Slote's 1989 example, he finds that 96% of the sample usage involves books that were last used or acquired since 1984, therefore, if he weeded out books that have not been used since 1983, he could still assure that 96% of usage in the library can be satisfied.

Prior to weeding, Slote suggests that the "elapsed time since installation" of the computer system is measured to assure that sample data is stable enough. Basically, he breaks this same circulation sample into three groups: (1) the ones that have been used before, (2) the ones that have not been used but were acquired *after* the installation of

the system, and (3) those that have not been used but were acquired *before* the time of installation. Slote recommended that the percentage of the last group must not exceed 15 percent.

6.3 Simon's Model and Its Applicability

In terms of the selection of resources, Simon's two assumptions as described in Section 1.3.2 can be restated to thusly:

(1) there is a constant probability, α, that the (t+1)-st circulation will be a new holding that has not been used in the first t circulations; and

(2) the probability that the (t+1)-st circulation is a holding that has been used n times is proportional to $n \cdot f(n,t)$, where f(n,t) is the number of distinct holdings that have circulated exactly n times each in the first t circulations of holdings.

Thus, the simulation mechanism is still valid in generating the usage patterns. Furthermore, the γ in the autoregressive model represents the aging/decay factor as described by researchers (Kent et al. 1979, Burrell 1980).

Since this chapter is a demonstration of applying our findings through analyzing the 80/20 rule using Simon's model, our usage pattern analysis is based on the methodology discussed in Section 4.5.

6.3.1 Library Data

We collected the usage information from the library information system of Sims Memorial Library of Southeastern Louisiana University, Hammond, LA. The following data fields were extracted from the database: (1) the item number assigned to each book,

(2) the number of its charge-outs (previous usage), (3) the year and date this book was last charged out, and (4) the status of the book (active or inactive). These raw data were then processed using a SPSS program to generate the $(n_i, f(n_i))$ dataset. We determined that there were 201,118 books in the library that are bar-coded and registered in the automated system, though the actual number could be somewhat higher. Excluding damaged and other inactive books, the total holdings available for circulation were 183,713. The system has been in place since Spring 1991, and we were informed that books which have not been bar-coded were not used during this period of time. Furthermore, the reserved books are required to be processed through the automated system, thus their usage is included in our study. The total number of transactions since the implementation of the automated system was 154,703.

Table 6.1 is the usage summary of the library data. The most active book was checked out 619 times (f(619) = 1) in the last 2½ years, and there were 31,113 books that have been checked out only one time each (f(1) = 31,113). The total number of books that have ever been used during this period was 61,606, or approximately 33.5% of the holdings that have been bar-coded. There are 103 clusters of books with different usages (m = 103), and 122,107 books had not been checked at all.

6.3.2 Estimating N

For our simulation purpose, we use N = 154,703, the total number of transactions since the automated system was first put in place in Spring 1991. Dividing N by 2.5, we obtain the annual average of 61,881 transactions. Suppose that we are interested in finding the usage pattern on an annual basis, then we would set N = 61,881.

Table 6.1: SLU Library Holding Usage Pattern

i	n_i	$f(n_i)$	i	n_i	$f(n_i)$	i	n_i	$f(n_i)$
1	1	31113	41	42	2	81	104	2
2	2	12913	42	43	3	82	106	2
3	3	6829	43	44	3	83	108	2
4	4	3769	44	45	3	84	113	1
5	5	2240	45	46	2	85	114	1
6	6	1431	46	47	1	86	115	1
7	7	896	47	48	2	87	117	3
8	8	665	48	49	3	88	118	2
9	9	441	49	51	4	89	120	1
10	10	308	50	52	3	90	123	1
11	11	207	51	53	2	91	126	1
12	12	151	52	56	2	92	131	1
13	13	122	53	57	3	93	146	1
14	14	78	54	59	4	94	157	1
15	15	62	55	60	1	95	161	1
16	16	51	56	62	2	96	165	1
17	17	30	57	63	2	97	188	1
18	18	30	58	64	2	98	224	1
19	19	16	59	65	1	99	236	1
20	20	14	60	66	1	100	305	1
21	21	14	61	69	2	101	311	1
22	22	14	62	70	1	102	590	1
23	23	14	63	71	3	103	619	1
24	24	12	64	72	1			
25	25	17	65	76	1			
26	26	4	66	77	1			
27	27	7	67	78	1			
28	28	8	68	79	1			
29	29	8	69	80	1			
30	30	9	70	84	3			
31	31	3	71	85	1			
32	32	1	72	86	1			
33	33	4	73	88	1			
34	34	6	74	90	1			
35	35	2	75	91	1			
36	36	1	76	94	1			
37	38	3	77	97	3			
38	39	3	78	99	1			
39	40	1	79	101	1			
40	41	7	80	103	1			

N = 154,703
T = 61,606
μ = 2.5112

6.3.3 Estimating α

Figure 6.1 is the 80/20 curve for the library data, and the curve tells us that the proper concentration measurement is really 70/30. That is, 70% of the total transactions involve approximately 30% of the bar-coded holdings. Our calculation shows that x_{m-1} = 0.49497 and θ_{m-1} = 0.79889, thus s_m = 0.398 for the SLU data — which, amazingly, is exactly the same as the ratio of the total number of holdings used (61,606) to the total transaction (154,703). Therefore, the α for our library data is determined to be approximately 0.40. For comparison purpose, from Table 2.1a we calculate Bradford's data to have s_m = 0.414, and the total number of authors is 41.5% of the total number of paper published. Similar phenomenon can be observed in Kendall's data as well. This outcome of T/N $\approx \alpha$ would be expected from simulation results: in Simon's model α is the probability of new entry, therefore the total number of books ever used should be the product of N and α. The amazing part is the exactness of outcomes shown even in empirical data regardless of different magnitudes and the selection processes.

6.3.4 Estimating γ

We ran several simulations with various γ but held N = 154,703 and α = 0.40 to generate 80/20 curves, and we compared the s_1 of the library data with the simulation results. We find that although s_1 is not as effective in determining γ as s_m is to α, through visual inspection of 80/20 curves it allows us to determine that γ falls in the neighborhood of 0.9995. If we use $\log(s_1)$ to estimate γ, then since $\log(s_1)$ = 2.6090 for γ = 1.0, 1.8379 for γ = 0.9995, and 2.3919 for SLU data; we interpolated γ to be 0.99986 in SLU data. However, when we compare the actual library data with

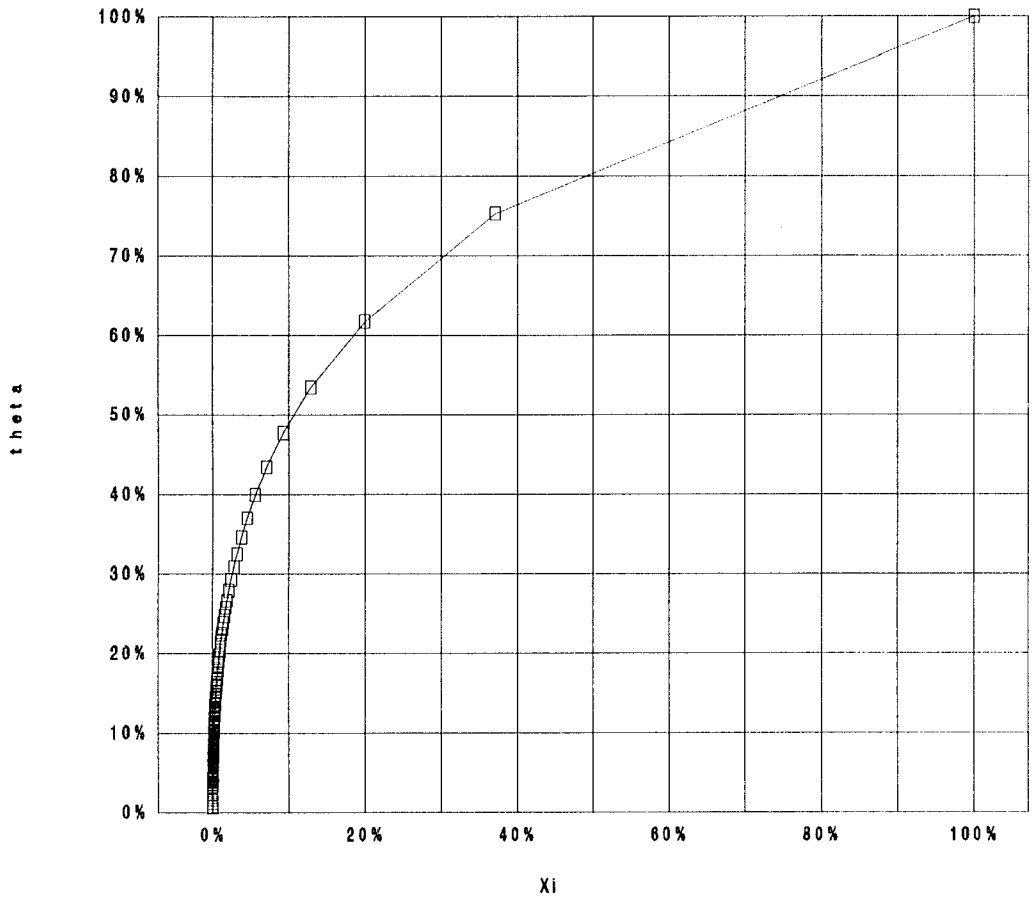


Figure 6.1: 80/20 Curve for Library Data

simulation result of $\gamma = 0.9995$, we find that the curve for SLU data falls below instead of above the simulated curve (Figure 6.2).

Table 6.2 is the simulated usage distribution. It shows that although it brings to question the predictive power of s_1 to γ, the general usage patterns between the simulated data and the actual data show remarkable resemblance, as demonstrated in Figure 6.2. Similarities are observed between Table 6.1 and Table 6.2 also. For instance, the average transaction per holding is 2.5418 in the simulated data (compare with 2.5112 of the actual data). On the other hand, even though the maximum index m in the simulated data is far less than the actual data, another simulation using $\alpha = 0.398$ and $\gamma = 0.99986$ yields $m = 76$, $n_m = 317$, $f(1) = 33{,}823$, and $T = 60632$ — a closer approximation. Therefore, Simon's model can simulate real usage patterns when the accuracy of parameter values is maximized.

6.4 Contributions to the Weeding Process

There are several contributions in supplementing Slote's method with Simon's model. Primarily, Simon's model generates quantifiable parameter values for decision makings, and it provides a sound theoretical foundation to the current method used.

A Quantifiable Method to Estimate the "Aging" Factor

We can use γ as a means to estimate the aging factor which in turn can be used to determine when a book should be weeded because of lack of use. At first glance, an aging factor of $\gamma = 0.99986$ for SLU data seems to contradict what has been recognized in the literature as use of library material decaying rapidly with time. However, when we consider that this γ represents the aging factor of *one transaction*, then we can

Table 6.2: Simulated Usage Distribution, $\alpha = 0.4$, $\gamma = 0.9995$, and N = 154,703

i	n_i	$f(n_i)$	i	n_i	$f(n_i)$
1	1	33788	41	45	2
2	2	10754	42	46	3
3	3	5094	43	50	1
4	4	3079	44	52	2
5	5	2188	45	56	1
6	6	1332	46	61	1
7	7	1012	47	62	1
8	8	715	48	63	1
9	9	610	49	68	1
10	10	476	50	90	1
11	11	311	51	129	1
12	12	236	52	138	1
13	13	248	53	175	1
14	14	186			
15	15	150			
16	16	116			
17	17	66			
18	18	81			
19	19	57			
20	20	55			
21	21	37			
22	22	50			
23	23	33			
24	24	27			
25	25	20			
26	26	16			
27	27	22			
28	28	11			
29	29	11			
30	30	14			
31	31	6			
32	32	4			
33	33	12			
34	34	2			
35	35	6			
36	36	3			
37	38	3			
38	39	6			
39	40	5			
40	41	4			

N = 154,703
T = 60,863
μ = 2.5418

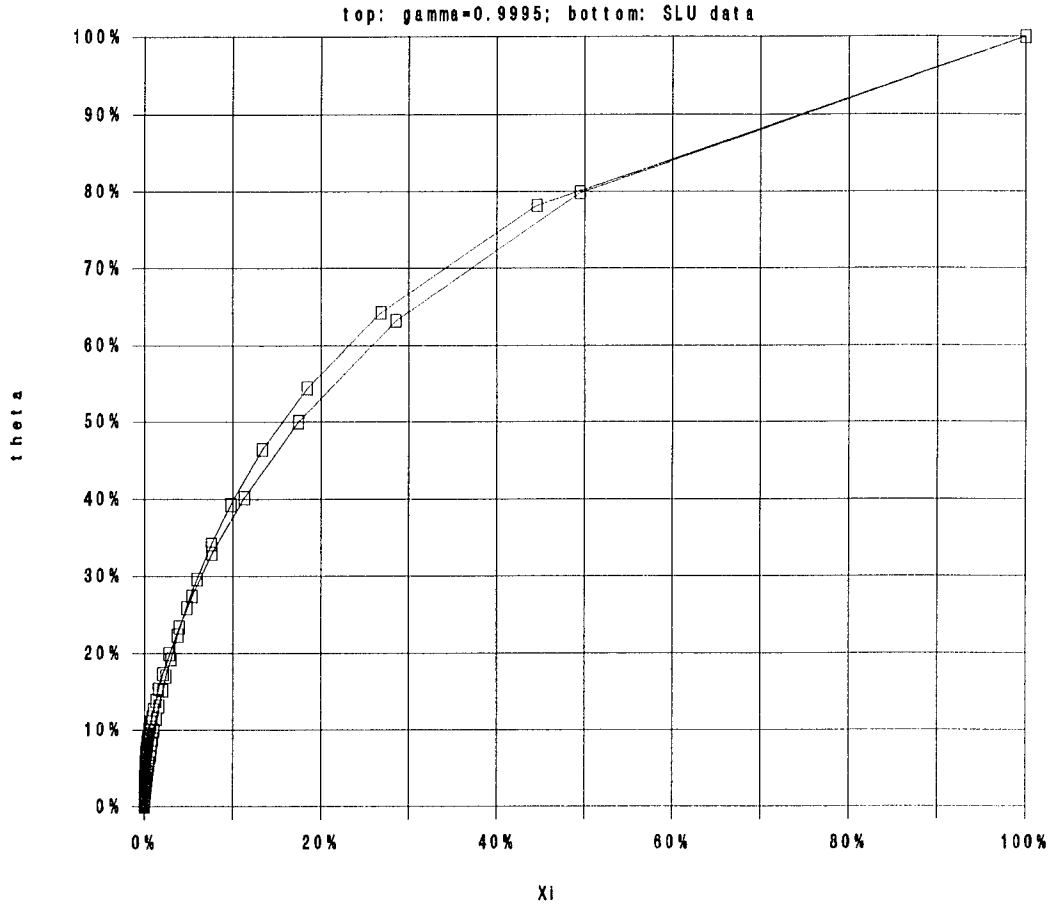


Figure 6.2: Simulation Results of the 80/20 Curve and Actual Library Data

calculate this aging factor for SLU library to be $(0.99986)^N$. Therefore, after one year (61,881 transactions) the probability of selection gained from the last selection only retains 0.00017 of its original weight. Although using this approach to determine weeding policy will require further calibration, results show that the aging is indeed rapid in library holdings, and this approach no longer requires the tedious and costly survey at the check-out counter.

A Quantifiable Method to Classify Libraries

The determinant of the adequate elapsed time since installation is really the probability that a new entry is not newly acquired, If we can calculate the α of several libraries we may find an α level that would indicate the stability of the sample data. Thus, instead of classifying libraries by descriptive characteristics such as "public libraries" or "medical libraries" or by nonproven attributes such as their different sizes, we now have a method of differentiate libraries according to their usage pattern; namely, their α and γ. Those libraries which are in the same category should be able to use similar weeding policies. This eliminates the need for surveying all libraries.

A Quantifiable Method of Estimating the Risk of Unavailability

Section 6.3.2 provides a method to estimate quickly the probability that a weeded book will be requested again. We can estimate α and multiply it by annual transaction to find the number of books that will be requested next year that have not previously being checked out. Therefore, we can conclude that there will be 24,752 books checked out for the first time next year. Suppose we store away the entire 122,107 books that

have not been used before, then there is a 20.3% chance that one of the weeded books would be needed next year.

Providing Theoretical Support for Small Sample Size

As illustrated in Bradford's 1934 data as well as SLU data, the proper concentration measurement should be about 70/30 rather than 80/20. More importantly, since the measurement is independent of N, as we have suggested in Chapter 4, Slote's method of using 400 sample consecutive transactions is supported by our model as being adequate.

Enhancing the Concept of "Keeping-Level"

Slote conscientiously separates transactions into groups according to their previous usage, and his "keeping-level" is essentially a measure of the current usage pattern. In other words, keeping-level reflects the end-result of the selection process up to this point. Suppose there are changes in the usage pattern, the library would not have a means of detecting such changes or anticipating future changes. By shifting the focus to the measurement of α and how it changes over time for different subclass of holdings, we have a better way to estimate future usages.

6.5 Possible Future Refinements

To verify that these parameters for a library are stable, we can repeat the estimation of α and τ once every six months. Over the period of time the characteristics of patrons may change, thus changing the usage pattern of the library. This longitudinal study may ensure that the library has an up-to-date policy to meet the changing demands of its users. We can also compare the α estimate of different time spans to see if α is

indeed a constant. It is possible that α is overstated in SLU data because of the newness of the automated system. It is possible that α becomes quite stable if the environment of the library holdings are stable, and will change when socio-economic conditions or the demography of its patrons change. We will need longitudinal study across more libraries to substantiate our hypotheses.

One of the problems in generating a usage pattern resembling SLU data is to generate an adequately high n_m. We need to note that 619 usages in 2½ years translated to over 20 usages each month. This is plausible in SLU data because it includes reserved books which may be checked out several times a day. Given that there may be several hundred students who are required to read the reserve material, we may want to separate these reserved holdings from the data and see if the usage pattern would change significantly. Unfortunately, under current automated system this process is still quite time consuming, and there may not be adequate records kept for all books that have ever been kept in reserve in all semesters.

CHAPTER 7

CONCLUSION

This dissertation, by using Chen's index approach, provides an analytical analysis for the 80/20 rule in an attempt to model information usage patterns. It also discusses other empirical phenomena that are closely associated with this skew distribution. In Chapter 3 we used the slope-distance pair to identify three important properties of the 80/20 curves without unreasonable assumptions and with fewer parameters than the traditional x (fraction of holdings) and θ (fraction of transactions) approach.

We then use Simon's three models of information usage and their generating mechanisms to produce the fundamental usage dataset (n_i, $f(n_i)$) — the frequency of usage and the number of information items which have been used that many times — to examine the effects of changing parameters of the 80/20 rule and other empirical phenomena. We find Simon's models to be simple yet robust in testing the extreme conditions. For instance, even though Simon's most complex model (the autoregressive model) has only two parameters — α (the probability of new entry) and γ (the decay/aging factor) — we were able to reproduce five of the six different classes of Bradford's curve, and the remaining one seem to be a combination of two classes. In fact, four of these five classes were generated using Simon's basic models which have only α as the parameter.

Through simulations we find the probability of new entry, α, to be the most influential factor in determining the shape of the 80/20 curve and other empirical phenomena. In general, simulation results show that with a low probability of new entry α, the usage becomes concentrated on a few items. On the other hand, lower γ (meaning more rapid decay) reduces the concentration of usage patterns. Since in Simon's autoregressive model he assumes one transaction takes place within a time unit, we can proxy the time span by the total number of iterations in our simulation model. Consequently, because the total number of usages has little effect on the simulation results once the system is stabilized (e.g., $N > 15{,}000$), we conclude that without changes in the fundamental factors such as α and γ, the time span has little effect on usage patterns except in the case that α is a decreasing function of time. Furthermore, assuming that the probability of new entry is at a constant rate, the 80/20 rule and most of other empirical laws hold at a very narrow range of $\alpha \approx 0.20$. When the decay factor, γ, is introduced, we find it to reduce the usage frequency of the most-used item and the number of items that are used only once. The end result is a more evenly distributed usage.

From examining Simon's assumptions, these results are plausible. Since we define higher α to mean more new entries, then given the same number of usages they will have to be distributed among more items, resulting in lower concentrations. Suppose $\alpha(R)$ is a decreasing function with respect to time, $R = 1,2,\ldots N$, then large N would produce smaller α in the long run thus increase the usage concentrations. On the other hand, the effect of "aging" or "decay" reduces usage concentrations by reducing

the dominance of the active items. A low γ means rapid reduction of the usage probability for those items which have not been used recently, thus providing an opportunity for inactive files to be selected in the future. When $\gamma = 1$, there is no aging factor, and therefore the probability of future usage of an item is strictly determined by its historical usage rate. Therefore, Simon's basic models are simply special cases of his autoregressive model.

The parameters in Simon's models affect usage patterns the following ways. Cluster 1, where $n_i = i$ consists of f(1) (the number of items that are used only once), is mostly influence by Simon's first assumption (new entry rate α). With a given α, the total number of different items to be used is fixed, therefore the distribution of usages is determined by the second assumption and the decay factor γ. Simon's second assumption reflects "success breeds success," while the γ in his autoregressive model allows the rising of new "success" by allowing a combination of repeated usage of new items and the decay of previously-active items to affect the usage probability structure. Simon's second assumption affects primarily the most-active item (whose usage frequency is n_m) by providing it higher probability of being used again. The decay factor γ, on the other hand, reduces n_m in time and thus decreases the concentration of the usage pattern over time.

Taking cue from our analytical finding (Equation 3.18), we find that α can be reliably estimated from an empirical dataset, and we applied this technique to the process of weeding library holdings. Since α determines f(1), we can use s_m of an empirical data to estimate its α. An alternative is to use T/N to estimate α, and these two approaches

yield similar results in empirical data testing. We can use this α and the given N of the empirical data to simulate the usage pattern of the dataset and to estimate its γ. The α and γ then can be used to show the characteristics of the empirical data. For instance, we estimated the usage pattern for SLU library holdings to have $\alpha \approx 0.398$ and $\gamma \approx 0.99986$. These parameter values provide a better identification of the usage pattern of the system than, say, the traditional 80/20 measure, since these indicators can help us assess the future usage patterns in a changing environment.

REFERENCES

Anderson, J.R., *The Adaptive Character of Thought*, Lawrence Erlbaum Associates, 1990.

Bach, L., "Models for the Location-Allocation-Problem in Urban and Regional Infrastructure Planning," *Journal of Interdisciplinary Modeling and Simulation*, Vol. 3, No. 3, 1980, pp. 285-326.

Boehm, B.W., "Improving Software Productivity," *Computer*, September 1987, pp. 43-57.

Booth, A.D., A law of occurrences for words of low frequency, *Information and control*, 10:386-393, 1967.

Bradford, S.C., "Sources of Information on Specific Subjects, *Engineering*, 137:85-86, 1934.

Buckland, M.K., *Book Availability and the Library Users*, New York: Pergamon; 1975.

Burrell, Q.L., "A Simple Stochastic Model for Library Loans," *Journal of Documentation*, Vol. 36, 1980, pp. 115-132.

Burrell, Q.L., "The 80/20 Rule: Library Lore or Statistical Law?" *Journal of Documentation*, 41:1, 24, 1985.

Chen, Y.S., "Analysis of Lotka's Law: the Simon-Yule Approach," *Information Processing & Management*, 25(5):527-544, 1989.

Chen, Y.S., "Zipf's Laws in Text Modeling," *International Journal of General Systems*, Vol. 15, pp. 233-252, 1989.

Chen, Y.S., "An Exponential Recurrence Distribution in the Simon-Yule Model of Text," *Cybernetics and Systems: An International Journal*, Vol. 19, pp. 521-545, 1988.

Chen, Y.S. and Leimkuhler, F.F., "Booth's Law of Word Frequency," *Journal of the American Society for Information Science*," 41(5):387-388, 1990.

Chen, Y.S. and Leimkuhler, F.F., "Analysis of Zipf's Law: an Index Approach," *Information Processing and Management*, 23(3):171-182, 1987.

Chen, Y.S., and F.F. Leimkuhler, "Bradford's Law: An Index Approach," *Scientometrics*, Vol. 11, No. 3-4, 1987, pp. 183-198.

Chen, Y.S., and F.F. Leimkuhler, "A Relationship Between Lotka's Law, Bradford's Law, and Zipf's Law," *Journal of the American Society for Information Science*, September 1986, pp. 307-314.

Chung, K.H., H.S. Pak, and R.A.K. Cox, "Patterns of Research Output in the Accounting Literature: A Study of the Bibliometric Distributions," *ABACUS*, Vol. 28, No. 2, 1992, pp. 168-185.

Chung, K.H., and R.A.K. Cox, "Patterns of Productivity in the Finance Literature: A Study of the Bibliometric Distributions," *Journal of Finance*, 65(1):301-309, 1990.

Clark, L., editor, *Guide to Review of Library Collections: Preservation, Storage, and Withdrawal*, American Library Association, 1991.

Coile, R.C., "Managing Technical Innovation," paper presented at the ORSA/TIMS Joint National Meeting, April 26, 1988, Washington, D.C.

Egghe, L, "On the 80/20 Rule," *Scientometrics*, 10(1-2):55-68, 1986

Evans, G.E., *Developing Library and Information Center Collections*, 2nd Ed., Libraries Unlimited, Inc., 1987.

Ijiri, Y., and H.A. Simon, *Skew Distributions and the Sizes of Business Firms*, North-Holland, 1977.

Kendall, M.G., "The Bibliography of Operational Research," *Operational Research Quarterly*, 11:31-36, 1960.

Kent, A. et al., *Use of Library Materials: The University of Pittsburgh Study*, New York: Marcel Dekker, 1979.

Korbert, N., "Managing Stock Dollars," *Purchasing*, July 24, 1980.

Kovacs, B., *The Decision-Making Process for Library Collections: Case Studies in Four Types of Libraries*, Greenwood Press, 1990.

Lancaster, F.W. and J.L. Lee, "Bibliometric Techniques Applied to Issues Management: A Case Study," *Journal of the American Society for Information Science*, 36(6):389-397, 1985.

Leimkuhler, F.F., "On Bibliometric Modeling," *Informetrics*, 97-104,1988.

Loomis, M.E.S., *Data Management and File Structures*, 2nd ed., Prentice-Hall, 1989, pp. 14-15.

Lotka, A.J., "The Frequency of Distribution of Scientific Productivity," *Journal of the Washington Academy of Science*, 16(12):317-323, 1926.

Merton (1973), cited in Creswell, J.W. *Faculty Research Performance: Lessons from the Sciences and the Social Sciences*, ASHE-ERIC, Higher Education Report No. 4. Washington D.C., 1985, p.

Morse, P.M. *Library Effectiveness: A System's Approach*, The M.I.T. Press, 1968.

Nash, K.S., "Keeping Pace with CASE Philosophy," *ComputerWorld*, August 17, 1992, pp. 81-83.

Neuts, M.F., "An Algorithmic Probabilist's Apology," in Gani, J. (ed.), *The Craft of Probabilistic Modelling*, Springer-Verlag, New York, 1986.

Neuts, M.F., "Computer Experimentation in Applied Probability," (working paper 86-030, Systems and Industrial Engineering Department, University of Arizona, 1986).

Pareto, V., *Manual of Political Economy*, 1909, English translation by A.M. Kelley, New York, 1971.

Sanders, R., "The Pareto Principle: Its Use and Abuse," *The Journal of Services Marketing*," Vol. 1, No. 2, Fall 1987, pp. 37-40.

Simon, H.A., *Models of My Life*, Basic Books (Harper Collins), 1991.

Simon, H.A., "On judging the plausibility of theories, in B. van Rootselaar and J. F. Staal (eds.)," *Logic, Methodology and Philosophy of Sciences*, Vol. III, Amsterdam: North-Holland, 1968.

Simon, H.A., "On a Class of Skew Distribution Function," *Biometrika*, 42:425-440, 1955.

Simon, H.A., and T.A. Van Wormer, "Some Monte Carlo Estimates of the Yule Distribution," *Behavior Science*, 1(8):203-210, 1963.

Slote, S.J., *Weeding Library Collections: Library Weeding Methods*, 3rd ed., Libraries Unlimited, Inc., Englewood, CO., 1989.

Stuart, J., "Mixing and Matching Pays off for Amoco Canada," *Computing Canada*, Vol. 18, No. 21, October 13, 1992, p. 59.

Tague, J., "What's the Use of Bibliometrics?" *Informetrics*, 1988, pp. 271-278.

Thiel, L.H., and H.S. Heaps, "Program Design for Retrospective Searches on Large Data Bases," *Information Storage and Retrieval*, 8(1972): 1-20, cited in H.S. Heaps, *Information Retrieval: Computational and Theoretical Aspects*, Academic Press, 1978, pp. 229-233.

White, H.D. and McCain, K.W., "Bibliometrics," *Annual Review of Information Science and Technology*, pp.119-186, Volume 24, 1989.

Zipf, G.K., *Human Behavior and the Principal of Least Effort*, Cambridge, MA, Addison-Wesley, 1949.

Zunde, P., "Empirical Laws and Theories of Information and Software Sciences," *Information Processing & Management*, Vol. 20, No. 1-2, 1984, pp. 5-18.

APPENDIX A: PROGRAM FOR BASIC MODELS

```
(*     This program generates information usage pattern              *)
(*     according to Simon's algorithm                                *)

Program Simon;
uses printer;

CONST
     N:longint = 20000;          (* N is total number of usage *)
     MAX = 20;                   (* MAX is m, the largest index *)

VAR
      f : ARRAY[1..MAX] OF INTEGER;
      (* f(1,k), only the last update kept *)

     a, b, bb, cumulant, alpha, r, mu : REAL;
     i, j, k, index, loop : INTEGER;

BEGIN
     alpha := 0.9;
     index := 0;
     f[1] := 3;                  (* initial condition f(1,0)=3 *)

      FOR i := 2 TO MAX DO f[i] := 0;       (* f(i,0)=0 for i>1 *)

     FOR k:= 4 TO N DO
     BEGIN
       a := RANDOM;
       writeln (k:5);

(* Note: replace alpha with A/log(k) for decreasing function model *)

       IF a <= alpha
       THEN     f[1] := f[1] + 1     (* Program step 1 *)
       ELSE
         BEGIN                       (* Program step 2 *)
           cumulant := 0;            (* cumulant = Σj·f(j,k-1) *)
           bb := RANDOM;
           b := 1 + bb*(k-2);        (* standardized b *)

           j := 0;
           REPEAT                    (* find the right f(n,k)  *)
             j := j + 1;
             cumulant := cumulant + j*f[j];
           UNTIL cumulant >= b;

           i := j;
           f[i] := f[i] - 1;
           f[i+1] := f[i+1] + 1;
         END; (* ELSE *)
     END;      (* k loop *)

     cumulant := 0;                  (* Calculate total usage count *)
     for i:=1 to MAX DO cumulant := cumulant + i*f[i];

(* output header, N, and α *)
```

```
     writeln (LST,'Word Count:',cumulant:10:0,'   alpha = ',alpha:4:2);
writeln;
     writeln (lst);    r :=0;

(* Print out the usage frequency distribution *)

     FOR i := 1 TO MAX DO
     BEGIN
       r := r + f[i];
       IF f[i]>0
       THEN
       BEGIN
         index := index + 1;
         WRITELN (LST, index:3,':','f(',i:4, ',',N:6,')',f[i]:10);
       END; (* IF-THEN-ELSE *)
     END;
     writeln(lst);
     writeln(lst, 'total f[i] =  ',r:10:0);
     mu := cumulant/r;  (* average usage *)
     writeln(lst);writeln(lst,'Average transactions per holding:',mu:10:4)
END.
```

APPENDIX B: PROGRAM FOR THE AUTOREGRESSIVE MODEL

```
(*************************************************************
 *                                                           *
 *       Simon's Third Model with Serial Correlation         *
 *                                                           *
 *              CONSTANT ENTRY RATE (ALPHA)                  *
 *                                                           *
 *           LAST UPDATE:   5/23/93   (MVS VERSION)          *
 *                                                           *
 *************************************************************)

PROGRAM SIMON3SC(OUTPUT);

CONST
     R   = 20000;                    (* NUMBER OF ITERATIONS EACH CYCLE *)
     RC = 1;                        (* NUMBER OF CYCLES EACH RUN  *)
     MAX = 40000;                    (* MAXIMUM NUMBER OF FIRMS EACLE *)
     MAXINDEX = 800;                     (* MAXIMUM RANK INDEX NUMBER *)

     ALPHA = 0.20;                       (* PROBABILITY OF NEW ENTRY  *)
     GAMMA = 0.85;                    (* GEOMETRIC RATE OF DIE OUT *)

     SEED1 = 123;
     SEED2 = 12345;
     SEED3 = 78901;
     SEED4 = 965431;

     SEED = SEED1;                     (* RANDOM # GENERATOR SEED    *)

TYPE
     AINT   = ARRAY(.1..MAX.) OF INTEGER;
     AREAL = ARRAY(.1..MAX.) OF REAL; (* FIRM SIZE AND WEIGHT TYPE *)
     AMAXI = ARRAY(.0..MAXINDEX.) OF INTEGER;

VAR
     s: AINT;                                (* CURRENT FIRM SIZE       *)
     w: AREAL;                               (* CURRENT WEIGHT OF FIRM  *)
     A, B, CUMULANT: REAL;
     nk: INTEGER;                            (* NO. OF FIRM CURRENTLY   *)
     aa: INTEGER;                            (* ASSETS ALLOCATED INITIALLY*)
     I, J, TEMP: INTEGER;                    (* LOOP COUNTER AND TEMP VAR *)
     c, ST, T: INTEGER;             (* CYCLE # AND ASSET RANGE ALLOCATED *)
     K: INTEGER;                             (* AGGREGATED ASSET SIZE *)
     MINREAL1: REAL;
     WEIGHTSUM: REAL;                        (* WEIGHT SUM OF ALL FIRMS *)


(* initialize # of firms with size and weight *)

procedure initialize;
var
   i: integer;
begin
   A := RANDOM(SEED);                      (* INITIALIZE RANDOM NUMBER *)
   MINREAL1 := 100*MINREAL;

   S(.1.) := 1;
```

```
   S(.2.) := 1;
   S(.3.) := 1;

   W(.1.) := 1;
   W(.2.) := 1;
   W(.3.) := 1;
   nk := 3;                          (* SET INITIAL FIRM DISTRIBUTION *)

   WEIGHTSUM := 0;
   aa := 0;
   for i:= 1 to nk do
   BEGIN
     AA := AA + S(.I.);              (* SUM INITIAL ASSETS ALLOCATED *)
     WEIGHTSUM := WEIGHTSUM + W(.I.)
   END
end;


(* CALCULATE THE CONCENTRATION MEASURE OF 80/20 RULE  *)

PROCEDURE GETA80(F, G: AMAXI; INDEXNOW: INTEGER);
VAR
  I: INTEGER;
  A80: REAL;

BEGIN
  A80 := 0;
  FOR I := 1 TO INDEXNOW DO
    A80 := A80 + (G(.I.) + G(.I-1.))*(F(.I.) - F(.I-1.));

  A80 := A80/(2*T*NK) - 0.5;
  WRITELN;
  WRITELN;
  WRITELN('    AREA OF CONCENTRATION OF 80/20 RULE ==== ', A80:6:4);
  WRITELN
END;


(* output current firm size and weight information in various format *)

procedure sortout(size:aint;  wt:areal; nof:integer);
var
   I, J, TEMP: INTEGER;
   rtemp: REAL;                    (* TEMPORARY VARIABLE FOR SWAPPING  *)
   fid: AINT;                      (* STORE FIRM IDENTIFICATION NUMBER *)
   CF, CG: AMAXI;
   INDEXNOW: INTEGER;

begin
   WRITELN;
   WRITELN;
   WRITELN('                    ***** FIRM SIZE AND WEIGHT LIST *****');
   write('- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -');
   WRITELN(' - - - - - - - - - -');
   WRITE('          IN ENTRY ORDER                    IN DECENDING');
   WRITELN(' ORDER OF SIZE');
   WRITE('- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -');
   WRITELN(' - - - - - - - - - -');
   FOR I := 1 TO NOF DO
     FID(.I.) := I;
                                   (* KEEP TRACK OF INITIAL FIRM ID *)
```

```
FOR I := 1 TO NOF-1 DO                (* SORT THE FIRM BASED ON SIZE *)
  FOR J := I+1 TO NOF DO              (* KEEP TRACK OF FIRM ID*)
    IF SIZE(.J.) > SIZE(.I.) THEN
    begin
      TEMP := SIZE(.I.);
      SIZE(.I.) := SIZE(.J.);
      SIZE(.J.) := TEMP;              (* SWAP FIRM SIZE *)

      RTEMP := WT(.I.);
      WT(.I.) := WT(.J.);
      WT(.J.) := RTEMP;               (* SWAP FIRM WEIGHT *)

      TEMP := FID(.I.);
      FID(.I.) := FID(.J.);
      FID(.J.) := TEMP
    END;                              (* KEEP TRACK OF FIRM ID     *)
  IF NOF>50 THEN J:=50
    ELSE J:= NOF;                     (* OUTPUT UPTO FIRST 50 FIRMS *)

FOR I := 1 TO J  DO                  (* PRINT OUT THE FINAL FIRM SIZE*)
BEGIN
  WRITE('S(',I:3,')=',S(.I.):5, ';    W(',I:3,') = ', W(.I.):7:4);
  WRITE('    S(',FID(.I.):5,') = ');
  WRITELN(SIZE(.I.):5, ';   W(',FID(.I.):5, ') = ', WT(.I.):7:4);
end;
PAGE;
WRITELN;
WRITELN;
WRITELN;
WRITE('RANK    SIZE     # OF FIRM     TOTAL WEIGHT    ');
WRITELN('INC. P. IF NOT NEW');
WRITE('----    ----     ---------     ------------    ');
WRITELN('-----------');
j := 1;                                 (* INITIAL RANK VALUE        *)
temp := 1;                              (* # OF FIRM OF SAME SIZE    *)
RTEMP := WT(.1.);                    (* SUM FIRMS' WEIGHT OF SAME SIZE *)
CG(.0.) := 0;
CF(.0.) := 0;
for i := 2 TO NOF DO                  (* WORK ON DESCENDING FIRM LIST *)
begin
  IF SIZE(.I-1.) > SIZE(.I.) THEN
  begin
    WRITE(J:3,':', SIZE(.I-1.):8,'    ', TEMP:8, RTEMP:16:4);
    IF RTEMP > (MINREAL*WEIGHTSUM) THEN
      WRITELN(RTEMP/WEIGHTSUM:16:4)
    ELSE
    BEGIN
      RTEMP := 0;
      WRITELN(RTEMP:16:4)          (* TO AVOID REAL TYPE UNDERFLOW *)
    END;
    CG(.J.) := CG(.J-1.) + SIZE(.I-1.)*TEMP;
    CF(.J.) := CF(.J-1.) + TEMP;     (* CUMULATIVE FREQUENCY, ETC.*)
    j := j+1;
    temp := 1;
    RTEMP := WT(.I.)
  end
  ELSE                                  (* FIRM OF EQUAL SIZE *)
  begin
    temp := temp+1;
    RTEMP := RTEMP+WT(.I.)
  end
```

```
    end;                                    (* MARGINAL CONDITION *)

    WRITE(J:3,':', SIZE(.NOF.):8,'    ', TEMP:8, RTEMP:16:4);
    WRITELN(RTEMP/WEIGHTSUM:16:4);
    CG(.J.) := CG(.J-1.) + SIZE(.NOF.)*TEMP;
    CF(.J.) := CF(.J-1.) + TEMP;
    GETA80(CF, CG, J)

end;


BEGIN   (* main *)

      initialize;
      PAGE;
      WRITELN;
      WRITELN;
      WRITELN;
      WRITE('         ');
      WRITELN('=============================================================');
      WRITE('         ');
      WRITELN('        ALPHA = ',ALPHA:3:2,'      GAMMA = ',GAMMA:5:4);
      WRITELN;
      WRITE('         ');
      WRITELN('    CYCLE SIZE = ',R:6, ';  # OF CYCLE EACH RUN = ',RC:2);
      WRITELN;
      WRITE('         ');
      WRITELN('               *** INITIAL CONDITIONS ***');
      WRITELN;
      WRITE('         ');
      WRITELN('           RANDOM # GENERATOR SEED = ', SEED:6);
      WRITELN;
      WRITE('         ');
      WRITE('    NUMBER OF FIRMS = ',NK:3, '   ASSETS ALLOCATED = ',AA:4);
      WRITELN;
      WRITE('         ');
      WRITELN('=============================================================');

      st :=  AA + 1;                (* ASSIGN INITIAL STARTING ASSET     *)
      T := 0;

      for c := 1 to rc do
      begin
        T := T + R;                 (* LOOP START/END # OF THIS CYCLE *)
        FOR k := st TO t DO
        BEGIN                       (* START ONE CYCLE OF A. ALLOCATION *)
          A := RANDOM(0);           (* GET RANDOM NUMBER IN (0,1)  *)
          IF a <= alpha
          THEN                                (* ALLOCATE TO A NEW FIRM   *)
          begin
            for i:=1 to nk do
              IF W(.I.) > MINREAL1 THEN     (* PREVENT DATA UNDERFLOW *)
                W(.I.) := W(.I.)*GAMMA;
                                    (* ADJUST WEIGHT OF EXISTING FIRMS *)
            nk := nk + 1;
            S(.NK.) := 1;
            W(.NK.) := 1            (* SIZE AND WEIGHT AT OUTSET  *)
          end
          ELSE
          BEGIN                     (* ALLOCATE TO AN EXISTING FIRM *)
            B := RANDOM(0)*WEIGHTSUM;     (* STANDARDIZED B       *)
```

```
            cumulant := 0;
            j := 0;
            REPEAT
              j := j + 1;
              CUMULANT := CUMULANT + W(.J.);
            UNTIL cumulant >= b;
            for i:=1 to nk do
              IF W(.I.) > MINREAL1 THEN    (* TO PREVENT UNDERFLOW *)
                W(.I.) := W(.I.)*GAMMA;    (* ADJUST ALL FIRM WEIGHTS *)

            W(.J.) := W(.J.) + 1;
            S(.J.) := S(.J.) + 1
          END;                             (* ALLOCATE TO EXISTING FIRM *)
        WEIGHTSUM := WEIGHTSUM*GAMMA + 1

        END;                               (* K LOOP: ONE CYCLE *)
        PAGE;
        WRITELN;
        WRITELN;
        WRITE('    ------------------------------------------------');
        WRITELN('--------');
        WRITE('      CURRENT PERIOD ENDING AFTER ASSETS ALLOCATED === ');
        WRITELN(T:6);
        writeln;
        WRITELN('      CURRENT TOTAL NUMBER OF FIRMS =', NK:5);
        WRITE('    ------------------------------------------------');
        WRITELN('--------');

        sortout(s, w, nk);

        ST := 1 + T;
      END          (* FOR LOOP *)
END.
```

VITA

The author immigrated to the United States in 1969. He graduated from Roosevelt High School in Seattle, WA, in 1972. Subsequently he returned to school in 1983 and received his B.A. in Psychology, A.S. in Computer Science, and M.B.A. from Southeastern Louisiana University, Hammond, LA, from 1983 to 1987. Upon receiving his M.B.A., he was appointed Instructor in Economics at SLU and worked closely with the Business Research department. During that period of time he conducted research in Louisiana economy, especially in labor market analysis. Later he was appointed Head of the Business Research Unit and Editor of *Southeastern Economic Outlook*, a triannual review published by the Business Research Unit. Currently the author teaches at the Department of Operations and Information Systems, School of Business Administration, Gonzaga University, Spokane, WA.

DOCTORAL EXAMINATION AND DISSERTATION REPORT

Candidate: Ping Pete Chong

Major Field: Business Administration (Quantitative Business Analysis)

Title of Dissertation: On Information Usage Modeling

Approved:

Ye-Sho Chen
Major Professor and Chairman

Daniel Fogel
Dean of the Graduate School

EXAMINING COMMITTEE:

J. C. Lin

James Pruett

Kwei Tang

Sumit Sarkar

David Blouin

Date of Examination:

August 2, 1993